THE BELL SYSTEM TECHNICAL JOURNAL

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

TABLE OF CONTENTS
AND
INDEX

VOLUME XXIX

1950

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

PRINTED IN U.S.A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXIX, 1950

Table of Contents

JANUARY, 1950

APRIL, 1950

JULY, 1950

OCTOBER, 1950

Index to Volume XXIX

H

L

M

N

O

P

Q

R

S

T

V

W

VOLUME XXIX JANUARY, 1950 NO. 1

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

50¢ per copy

$1.50 per Year

THE BELL SYSTEM TECHNICAL JOURNAL

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

SUBSCRIPTIONS

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

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XXIX *January, 1950* *No. 1*

Traveling-Wave Tubes

By J. R. PIERCE

The following material on traveling-wave tubes is taken from a book which will be published by Van Nostrand in September, 1950. Substantially the entire contents of the book will be published in this and the three succeeding issues of the Bell System Technical Journal.

This material will cover in detail the theory of traveling-wave amplifiers. In addition, brief discussions of magnetron amplifiers and double-stream amplifiers are included. Experimental material is drawn on in a general way only, as indicating the range of validity of the theoretical treatments.

The material deals only with the high-frequency electronic and circuit behavior of tubes. Such matters as matching into circuits are not considered; neither are problems of beam formation and electron focusing, which have been dealt with elsewhere.[1]

The material opens with the presentation of a simplified theory of the traveling-wave tube. A discussion of circuits follows, including helix calculations, a treatment of filter-type circuits, some general circuit considerations which show that gain will be highest for low group velocities and low stored energies, and a justification of a simple transmission line treatment of circuits by means of an expansion in terms of the normal modes of propagation of a circuit. Then a detailed analysis of overall electronic and circuit behavior is made, including a discussion of various electronic and circuit waves, the fitting of boundary conditions to obtain overall gain, noise figure calculations, transverse motions of electrons and field solutions appropriate to broad electron streams. Short treatments of the magnetron amplifier and the double-stream amplifier follow.

[1] For instance, "Theory and Design of Electron Beams," J. R. Pierce, Van Nostrand, 1949.

CHAPTER I

INTRODUCTION

ASTRONOMERS are interested in stars and galaxies, physicists in atoms and crystals, and biologists in cells and tissues because these are natural objects which are always with us and which we must understand. The traveling-wave tube is a constructed complication, and it can be of interest only when and as long as it successfully competes with older and newer microwave devices. In this relative sense, it is successful and hence important.

This does not mean that the traveling-wave tube is better than other microwave tubes in all respects. As yet it is somewhat inefficient compared with most magnetrons and even with some klystrons, although efficiencies of over 10 per cent have been attained. It seems reasonable that the efficiency of traveling-wave tubes will improve with time, and a related device, the magnetron amplifier, promises high efficiencies. Still, efficiency is not the chief merit of the traveling-wave tube.

Nor is gain, although the traveling-wave tubes have been built with gains of over 30 db, gains which are rivaled only by the newer double-stream amplifier and perhaps by multi-resonator klystrons.

In noise figure the traveling-wave tube appears to be superior to other microwave devices, and noise figures of around 12 db have been reported. This is certainly a very important point in its favor.

Structurally, the traveling-wave tube is simple, and this too is important. Simplicity of structure has made it possible to build successful amplifiers for frequencies as high as 48,000 megacycles (6.25 mm). When we consider that successful traveling-wave tubes have been built for 200 mc, we realize that the traveling-wave amplifier covers an enormous range of frequencies.

The really vital feature of the traveling-wave tube, however, the new feature which makes it different from and superior to earlier devices, is its tremendous bandwidth.

It is comparatively easy to build tubes with a 20 per cent bandwidth at 4,000 mc, that is, with a bandwidth of 800 mc, and L. M. Field has reported a bandwidth of 3 to 1 extending from 350 mc to 1,050 mc. There seems no reason why even broader bandwidths should not be attained.

As it happens, there is a current need for more bandwidth in the general field of communication. For one thing, the rate of transmission of intelli-

gence by telegraph, by telephone or by facsimile is directly proportional to bandwidth; and, with an increase in communication in all of these fields, more bandwidth is needed.

Further, new services require much more bandwidth than old services. A bandwidth of 4,000 cycles suffices for a telephone conversation. A bandwidth of 15,000 cycles is required for a very-high-fidelity program circuit. A single black-and-white television channel occupies a bandwidth of about 4 mc, or approximately a thousand times the bandwidth required for telephony.

Beyond these requirements for greater bandwidth to transmit greater amounts of intelligence and to provide new types of service, there is currently a third need for more bandwidth. In FM broadcasting, a radio frequency bandwidth of 150 kc is used in transmitting a 15 kc audio channel. This ten-fold increase in bandwidth does not represent a waste of frequency space, because by using the extra bandwidth a considerable immunity to noise and interference is achieved. Other attractive types of modulation, such as PCM (pulse code modulation) also make use of wide bandwidths in overcoming distortion, noise and interference.

At present, the media of communication which have been used in the past are becoming increasingly crowded. With a bandwidth of about 3 mc, approximately 600 telephone channels can be transmitted on a single coaxial cable. It is very hard to make amplifiers which have the high quality necessary for single sideband transmission with bandwidths more than a few times broader than this. In television there are a number of channels suitable for local broadcasting in the range around 100 mc, and amplifiers sufficiently broad and of sufficiently good quality to amplify a single television channel for a small number of times are available. It is clear, however, that at these lower frequencies it would be very difficult to provide a number of long-haul television channels and to increase telephone and other services substantially.

Fortunately, the microwave spectrum, wh'ch has been exploited increasingly since the war, provides a great deal of new frequency space. For instance, the entire broadcast band, which is about 1 mc wide, is not sufficient for one television signal. The small part of the microwave spectrum in the wavelength range from 6 to $7\frac{1}{2}$ cm has a frequency range of 1,000 mc, which is sufficient to transmit many simultaneous television channels, even when broad-band methods such as FM or PCM are used.

In order fully to exploit the microwave spectrum, it is desirable to have amplifiers with bandwidths commensurate with the frequency space available. This is partly because one wishes to send a great deal of information in the microwave range: a great many telephone channels and a substantial number of television channels. There is another reason why very broad

bands are needed in the microwave range. In providing an integrated nationwide communication service, it is necessary for the signals to be amplified by many repeaters. Amplification of the single-sideband type of signal used in coaxial systems, or even amplification of amplitude modulated signals, requires a freedom from distortion in amplifiers which it seems almost impossible to attain at microwave frequencies, and a freedom from interfering signals which it will be very difficult to attain. For these reasons, it seems almost essential to rely on methods of modulation which use a large bandwidth in order to overcome both amplifier distortion and also interference.

Many microwave amplifiers are inferior in bandwidth to amplifiers available at lower frequencies. Klystrons give perhaps a little less bandwidth than good low-frequency pentodes. The type 416A triode, recently developed at Bell Telephone Laboratories, gives bandwidths in the 4,000 mc range somewhat larger than those attainable at lower frequencies. Both the klystron and the triode have, however, the same fundamental limitation as do other conventional tubes. As the band is broadened at any frequency, the gain is necessarily decreased, and for a given tube there is a bandwidth beyond which no gain is available. This is so because the signal must be applied by means of some sort of resonant circuit across a capacitance at the input of the tube.

In the traveling-wave tube, this limitation is overcome completely. There is no input capacitance nor any resonant circuit. The tube is a smooth transmission line with a negative attenuation in the forward direction and a positive attenuation in the backward direction. The bandwidth can be limited by transducers connecting the circuit of the tube to the source and the load, but the bandwidth of such transducers can be made very great. The tube itself has a gradual change of gain with frequency, and we have seen that this allows a bandwidth of three times and perhaps more. This means that bandwidths of more than 1,000 mc are available in the microwave range. Such bandwidths are indeed so great that at present we have no means for fully exploiting them.

In all, the traveling-wave tube compares favorably with other microwave devices in gain, in noise figure, in simplicity of construction and in frequency range. While it is not as good as the magnetron in efficiency, reasonable efficiencies can be attained and greater efficiencies are to be expected. Finally, it does provide amplification over a bandwidth commensurate with the frequency space available at microwaves.

The purpose of this book is to collect and present theoretical material which will be useful to those who want to know about, to design or to do research on traveling-wave tubes. Some of this material has appeared in print. Other parts of the material are new. The old material and the new material have been given a common notation.

The material covers the radio-frequency aspects of the electronic behavior of the tube and its internal circuit behavior. Matters such as matching into and out of the slow-wave structures which are described are not considered. Neither are problems of producing and focusing electron beams, which have been discussed elsewhere,[1] nor are those of mechanical structure nor of heat dissipation.

In the field covered, an effort has been made to select material of practical value, and to present it as understandably as possible. References to various publications cover some of the finer points. The book refers to experimental data only incidentally in making general evaluations of theoretical results.

To try to present the theory of the traveling-wave tube is difficult without some reference to the overall picture which the theory is supposed to give. One feels in the position of lifting himself by his bootstraps. For this reason the following chapter gives a brief general description of the traveling-wave tube and a brief and specialized analysis of its operation. This chapter is intended to give the reader some insight into the nature of the problems which are to be met. In Chapters III through VI, slow-wave circuits are discussed to give a qualitative and quantitative idea of their nature and limitations. Then, simplified equations for the overall behavior of the tube are introduced and solved, and matters such as overall gain, insertion of loss, a-c space-charge effects, noise figure, field analysis of operation and transverse field operation are considered. A brief discussion of power output is given.

Two final chapters discuss briefly two closely related types of tube; the traveling-wave magnetron amplifier and the double-stream amplifier.

[1] loc. cit.

CHAPTER II

SIMPLE THEORY OF TRAVELING-WAVE TUBE GAIN

Synopsis of Chapter

IT IS difficult to describe general circuit or electronic features of traveling-wave tubes without some picture of a traveling-wave tube and traveling-wave gain. In this chapter a typical tube is described, and a simple theoretical treatment is carried far enough to describe traveling-wave gain in terms of an increasing electromagnetic and space-charge wave and to express the rate of increase in terms of electronic and circuit parameters.

In particular, Fig. 2.1 shows a typical traveling-wave tube. The parts of this (or of any other traveling-wave tube) which are discussed are the electron beam and the slow-wave circuit, represented in Fig. 2.2 by an electron beam and a helix.

In order to derive equations covering this portion of the tube, the properties of the helix are simulated by the simple delay line or network of Fig. 2.3, and ordinary network equations are applied. The electrons are assumed to flow very close to the line, so that all displacement current due to the presence of electrons flows directly into the line as an impressed current

For small signals a wave-type solution of the equations is known to exist, in which all a-c electronic and circuit quantities vary with time and distance as $\exp(j\omega t - \Gamma z)$. Thus, it is possible to assume this from the start.

On this basis the excitation of the circuit by a beam current of this form is evaluated (equation (2.10)). Conversely, the beam current due to a circuit voltage of this form is calculated (equation (2.22)). If these are to be consistent, the propagation constant Γ must satisfy a combined equation (2.23).

The equation for the propagation constant is of the fourth degree in Γ, so that any disturbance of the circuit and electron stream may be expressed as a sum of four waves.

Because some quantities are in practical cases small compared with others, it is possible to obtain good values of the roots by making an approximation. This reduces the equation to the third degree. The solutions are expressed in the form

$$-\Gamma = -j\beta_e + \beta_e C\delta$$

Here β_e is a phase constant corresponding to the electron velocity (2.16) and C is a gain parameter depending on circuit and beam impedance (2.43). A solution of the equation for the case of an electron speed equal to the speed of the undisturbed wave yields 3 values of δ which are shown in Fig. 2.4. These represent an increasing, a decreasing and an unattenuated wave. The increasing wave is of course responsible for the gain of the tube. A different approximation yields the missing backward unattenuated wave (2.32).

The characteristic impedance of the forward waves is expressed in terms of β_e, C, and δ (2.36) and is found to differ little from the impedance in the absence of electrons.

The gain of the increasing wave is expressed in terms of C and the length of the tube in wavelengths, N

$$G = 47.3\,CN \text{ db} \tag{2.37}$$

It will be shown later that the gain of the tube can be expressed approximately as the sum of the gain of the increasing wave plus a constant to take into account the setting up of the increasing wave, or the boundary conditions (2.39).

Finally, the important gain parameter C is discussed. The circuit part of this parameter is measured by the cube root of an impedance, $(E^2/\beta^2 P)^{1/3}$, which relates the peak field E acting on the electrons, the phase constant $\beta = \omega/v$, and the power flow. $(E^2/\beta^2 P)^{1/3}$ is a measure of circuit goodness as far as gain is concerned.

We should note also that a desirable circuit property is constancy of phase velocity with frequency, for the electron velocity must be near to the circuit phase velocity to produce gain.

Evaluation of the effects of attenuation, of varying the electron velocity and many other matters are treated in later chapters.

2.1 Description of a Traveling-Wave Tube

Figure 2.1 shows a typical traveling-wave tube such as may be used at frequencies around 4,000 megacycles. Such a tube may operate with a cathode current of around 10 ma and a beam voltage of around 1500 volts. There are two essential parts of a traveling-wave amplifier; one is the helix, which merely serves as a means for producing a slow electromagnetic wave with a longitudinal electric field; and the other is the electron flow. At the input the wave is transferred from a wave guide to the helix by means of a short antenna and similarly at the output the wave is transferred from the helix to a short antenna from which it is radiated into the output wave guide. The wave travels along the wire of the helix with approximately the speed of light. For operation at 1500 volts, corresponding to about $\frac{1}{13}$ the

speed of light, the wire in the helix will be about thirteen times as long as the axial length of the helix, giving a wave velocity of about $\frac{1}{13}$ the speed of light along the axis of the helix. A longitudinal magnetic focusing field of a few hundred gauss may be used to confine the electron beam and enable it to pass completely through the helix, which for 4000 megacycle operation may be around a foot long.

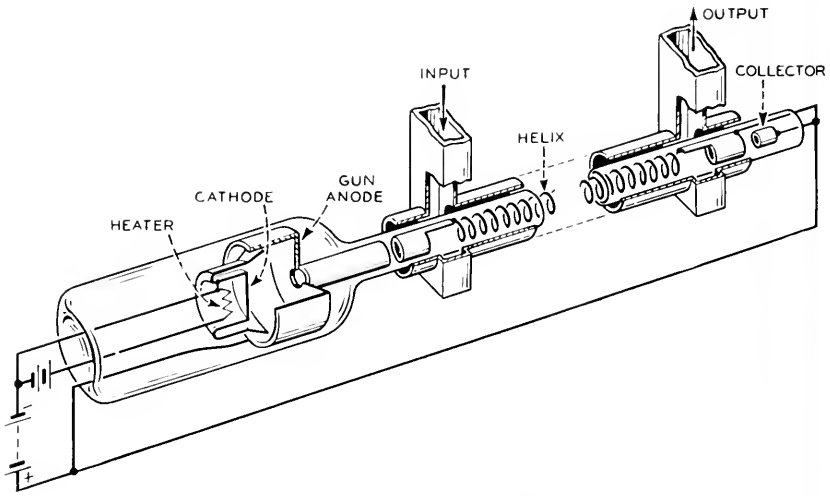

Fig. 2.1—Schematic of the traveling-wave amplifier.

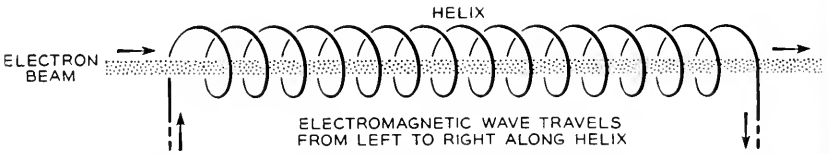

Fig. 2.2—Portion of the traveling-wave amplifier pertaining to electronic interaction with radio-frequency fields and radio-frequency gain.

In analyzing the operation of the traveling-wave tube, it is necessary to focus our attention merely on the two essential parts shown in Fig. 2.2, the circuit (helix) and the electron stream.

2.2 The Type of Analysis Used

A mathematical treatment of the traveling-wave tube is very important, not so much to give an exact numerical prediction of operation as to give a picture of the operation and to enable one to predict at least qualitatively the effect of various physical variations or features. It is unlikely that all of

the phenomena in a traveling-wave tube can be satisfactorily described in a theory which is simple enough to yield useful results. Most analyses, for instance, deal only with the small-signal or linear theory of the traveling-wave tube. The distribution of current in the electron beam can have an important influence on operation, and yet in an experimental tube it is often difficult to tell just what this distribution is. Even the more elaborate analyses of linear behavior assume a constant current density across the beam. Similarly, in most practical traveling-wave tubes, a certain fraction of the current is lost on the helix and yet this is not taken into account in the usual theories.

It has been suggested that an absolutely complete theory of the traveling-wave tube is almost out of the question. The attack which seems likely to yield the best numerical results is that of writing the appropriate partial differential equations for the disturbance in the electron stream inside the helix and outside of the helix. This attack has been used by Chu and Jackson[2] and by Rydbeck.[3] While it enables one to evaluate certain quantities which can only be estimated in a simpler theory, the general results do not differ qualitatively and are in fair quantitative agreement with those which are derived here by a simpler theory.

In the analysis chosen here, a number of approximations are made at the very beginning. This not only simplifies the mathematics but it cuts down the number of parameters involved and gives to these parameters a simple physical meaning. In terms of the parameters of this simple theory, a great many interesting problems concerning noise, attenuation and various boundary conditions can be worked out. With a more complicated theory, the working out of each of these problems would constitute essentially a new problem rather than a mere application of various formulae.

There are certain consequences of a more general treatment of a traveling-wave tube which are not apparent in the simple theory presented here. Some of these matters will be discussed in Chapters XII, XIII and XIV.

The theory presented here is a small signal theory. This means that the equations governing electron flow have been linearized by neglecting certain quantities which become negligible when the signals are small. This results in a wave-type solution. Besides the small signal limitation of the analyses presented here, the chief simplifying assumption which has been made is that all the electrons in the electron flow are acted on by the same a-c field, or at least by known fields. The electrons will be acted on by essentially the same field when the diameter of the electron beam is small enough or when

[2] L. J. Chu and J. D. Jackson, "Field Theory of Traveling-Wave Tubes," *Proc. I. R. E.*, Vol. 36, pp. 853–863, July 1948.

[3] Olof E. H. Rydbeck, "The Theory of the Traveling-Wave Tube," *Ericsson Technics*, No. 46, 1948.

the electrons form a hollow cylindrical beam in an axially symmetrical circuit, a case of some practical importance.

Besides these assumptions, it is assumed in this section that the electrons are displaced by the a-c field in the axial direction only. This may be approximately true in many cases and is essentially so when a strong magnetic focusing field is used. The effects of transverse motion will be discussed in Chapter XIII.

In this chapter an approximate relation suitable for electron speeds small compared to the velocity of light is used in computing interaction between electrons and the circuit.

A more general relation between impressed current and circuit field, valid for faster waves, will be given in Chapter VI. Non-relativistic equations of motion will, however, be used throughout the book. With whatever speed the waves travel, it will be assumed that the electron speed is always small compared with the speed of light.

We consider here the interaction between an electric circuit capable of propagating a slow electromagnetic wave and a stream of electrons. We can consider that the signal current in the circuit is the result of the disturbed electron stream acting on the circuit and we can consider that the disturbance on the electron stream is the result of the fields of the circuit acting on the electrons. Thus the problem naturally divides itself into two parts.

2.3 The Field Caused by an Impressed Current

We will first consider the problem of the disturbance produced in the circuit by a bunched electron stream. In considering this problem in this section in a manner valid for slow waves and small electron velocities, we will use the picture in Fig. 2.3. Here we have a circuit or network with uniformly

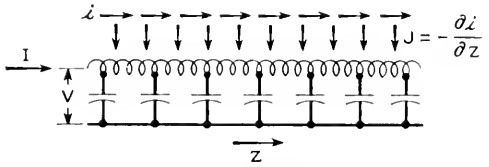

Fig. 2.3 —Equivalent circuit of a traveling-wave tube. The distributed inductance and capacitance are chosen to match the phase velocity and field strength of the field acting on the electrons. The impressed current due to the electrons is $-\partial i/\partial z$, where i is the electron convection current.

distributed series inductance and shunt capacitance and with current I and voltage V. The circuit extends infinitely in the z direction. An electron convection current i flows along very close to the circuit. The sum of the displacement and convection current into any little volume of the electron beam must be zero. Because the convection current varies with distance in

the direction of flow, there will be a displacement current J amperes per meter impressed on the transmission circuit. We will assume that the electron beam is very narrow and very close to the circuit, so that the displacement current along the stream is negligible compared with that from the stream to the circuit. In this case the displacement current to the circuit will be given by the rate of change of the convection current with distance.

If the convection current i and the impressed current J are sinusoidal with time, the equations for the network shown in Fig. 2.3 are

$$\frac{\partial I}{\partial z} = -jBV + J \tag{2.1}$$

$$\frac{\partial V}{\partial z} = -jXI \tag{2.2}$$

Here I and V are the current and the voltage in the line, B and X are the shunt susceptance and series reactance per unit length and J is the impressed current per unit length.

It may be objected that these "network" equations are not valid for a transmission circuit operating at high frequencies. Certainly, the electric field in such a circuit cannot be described by a scalar electric potential. We can, however, choose BX so that the phase velocity of the circuit of Fig. 2.3 is the same as that for a particular traveling-wave tube. We can further choose X/B so that, for unit power flow, the longitudinal field acting on the electrons according to Fig. 2.3, that is, $-\partial V/\partial z$, is equal to the true field for a particular circuit. This lends a plausibility to the use of (2.1) and (2.2). The fact that results based on these equations are actually a good approximation for phase velocities small compared with the velocity of light is established in Chapter VI.

We will be interested in cases in which all quantities vary with distance as $\exp(-\Gamma z)$. Under these circumstances, we can replace differentiation with respect to z by multiplication by $-\Gamma$. The impressed current per unit length is given by

$$J = -\frac{\partial i}{\partial z} = \Gamma i \tag{2.3}$$

Equations (2.1) and (2.2) become

$$-\Gamma I = -jBV + \Gamma i \tag{2.4}$$

$$-\Gamma V = -jXI \tag{2.5}$$

If we eliminate I, we obtain

$$V(\Gamma^2 + BX) = -j\Gamma Xi \tag{2.6}$$

Now, if there were no impressed current, the righthand side of (2.6) would be zero and (2.6) would be the usual transmission-line equation. In this case, Γ assumes a value Γ_1, the natural propagation constant of the line, which is given by

$$\Gamma_1 = j\sqrt{BX} \tag{2.7}$$

The forward wave on the line varies with distance as $\exp(-\Gamma_1 z)$ and the backward wave as $\exp(+\Gamma_1 z)$.

Another important property of the line itself is the characteristic impedance K, which is given by

$$K = \sqrt{X/B} \tag{2.8}$$

We can express the series reactance X in terms of Γ_1 and K

$$X = -jK\Gamma_1 \tag{2.9}$$

Here the sign has been chosen to assure that X is positive with the sign given in (2.7). In terms of Γ_1 and K, (2.6) may be written

$$V = \frac{-\Gamma\Gamma_1 K i}{(\Gamma^2 - \Gamma_1^2)} \tag{2.10}$$

In (2.10), the convection current i is assumed to vary sinusoidally with time and as $\exp(-\Gamma z)$ with distance. This current will produce the voltage V in the line. The voltage of the line given by (2.10) also varies sinusoidally with time and as $\exp(-\Gamma z)$ with distance.

2.4 Convection Current Produced by the Field

The other part of the problem is to find the disturbance produced on the electron stream by the fields of the line. In this analysis we will use the quantities listed below, all expressed in M.K.S. units.[4]

η—charge-to-mass ratio of electrons
$\eta = 1.759 \times 10^{11}$ coulomb/kg
u_0—average velocity of electrons
V_0—voltage by which electrons are accelerated to give them the velocity u_0. $u_0 = \sqrt{2\eta V_0}$
I_0—average electron convection current
ρ_0—average charge per unit length
$\rho_0 = -I_0/u_0$
v—a-c component of velocity
ρ—a-c component of linear charge density
i—a-c component of electron convection current

[4] Various physical constants are listed in Appendix 1.

The quantities v, ρ, and i are assumed to vary with time and distance as $\exp(j\omega t - \Gamma z)$.

One equation we have concerning the motion of the electrons is that the time rate of change of velocity is equal to the charge-to-mass ratio times the electric gradient.

$$\frac{d(u_0 + v)}{dt} = \eta \frac{\partial V}{\partial z} \tag{2.11}$$

In (2.11) the derivative represents the change of velocity observed in following an individual electron. There is, of course, no change in the average velocity u_0. The change in the a-c component of velocity may be expressed in terms of partial derivatives, $\frac{\partial v}{\partial t}$, which is the rate of change with time of the velocity of electrons passing a given point, and $\frac{\partial v}{\partial z}$, which is variation of electron velocity with distance at a fixed time.

$$\frac{dv}{dt} = \frac{\partial v}{\partial t} + \frac{\partial v}{\partial z}\frac{dz}{dt} = \eta \frac{\partial V}{\partial z} \tag{2.12}$$

Equation (2.12) may be rewritten

$$\frac{\partial v}{\partial t} + \frac{\partial v}{\partial z}(u_0 + v) = \eta \frac{\partial V}{\partial z} \tag{2.13}$$

Now it will be assumed that the a-c velocity v is very small compared with the average velocity u_0, and v will be neglected in the parentheses. The reason for doing this is to obtain differential equations which are linear, that is, in which products of a-c terms do not appear. Such linear equations necessarily give a wave type of variation with time and distance, such as we have assumed. The justification for neglecting products of a-c terms is that we are interested in the behavior of traveling-wave tubes at small signal levels, and that it is very difficult to handle the non-linear equations. When we have linearized (2.13) we may replace the differentiation with a respect to time by multiplication by $j\omega$ and differentiation with respect to distance by multiplication by $-\Gamma$ and obtain

$$(j\omega - u_0\Gamma)v = -\eta\Gamma V \tag{2.14}$$

We can solve (2.14) for the a-c velocity and obtain

$$v = \frac{-\eta\Gamma V}{u_0(j\beta_e - \Gamma)} \tag{2.15}$$

Where

$$\beta_e = \omega/u_0 \tag{2.16}$$

We may think of β_e as the phase constant of a disturbance traveling with the electron velocity.

We have another equation to work with, a relation which is sometimes called the equation of continuity and sometimes the equation of conservation of charge. If the convection current changes with distance, charge must accumulate or decrease in any small elementary distance, and we see that in one dimension the relation obeyed must be

$$\frac{\partial i}{\partial z} = -\frac{\partial \rho}{\partial t} \tag{2.17}$$

Again we may proceed as before and solve for the a-c charge density ρ

$$-\Gamma i = -j\omega\rho$$

$$\rho = \frac{-j\Gamma i}{\omega} \tag{2.18}$$

The total convection current is the total velocity times the total charge density

$$-I_0 + i = (u_0 + v)(\rho_0 + \rho) \tag{2.19}$$

Again we will linearize this equation by neglecting products of a-c quantities in comparison with products of a-c quantities and a d-c quantity. This gives us

$$i = \rho_0 v + u_0 \rho \tag{2.20}$$

We can now substitute the value ρ obtained from (2.18) into (2.20) and solve for the convection current in terms of the velocity, obtaining

$$i = \frac{j\beta_e \rho_0 v}{(j\beta_e - \Gamma)} \tag{2.21}$$

Using (2.15) which gives the velocity in terms of the voltage, we obtain the convection current in terms of the voltage

$$i = \frac{jI_0 \beta_e \Gamma V}{2V_0(j\beta_e - \Gamma)^2} \tag{2.22}$$

2.5 Overall Circuit and Electronic Equation

In (2.22) we have the convection current in terms of the voltage. In (2.10) we have the voltage in terms of the convection current. Any value of Γ for which both of these equations are satisfied represents a natural mode of

propagation along the circuit and the electron stream. When we combine (2.22) and (2.10) we obtain as the equation which Γ must satisfy:

$$1 = \frac{jKI_0\beta_e\Gamma^2\Gamma_1}{2V_0(\Gamma_1^2 - \Gamma^2)(j\beta_e - \Gamma)^2} \tag{2.23}$$

Equation (2.23) applies for any electron velocity, specified by β_e, and any wave velocity and attenuation, specified by the imaginary and real parts of the circuit propagation constant Γ_1. Equation (2.23) is of the fourth degree. This means that it will yield four values of Γ which represent four natural modes of propagation along the electron stream and the circuit. The circuit alone would have two modes of propagation, and this is consistent with the fact that the voltages at the two ends can be specified independently, and hence two boundary conditions must be satisfied. Four boundary conditions must be satisfied with the combination of circuit and electron stream. These may be taken as the voltages at the two ends of the helix and the a-c velocity and a-c convection current of the electron stream at the point where the electrons are injected. The four modes of propagation or the waves given by (2.23) enable us to satisfy these boundary conditions.

We are particularly interested in a wave in the direction of electron flow which has about the electron speed and which will account for the observed gain of the traveling-wave tube. Let us assume that the electron speed is made equal to the speed of the wave in the absence of electrons, so that

$$-\Gamma_1 = -j\beta_e \tag{2.24}$$

As we are looking for a wave with about the electron speed, we will assume that the propagation constant differs from β_e by a small amount ξ, so that

$$\begin{aligned} -\Gamma &= -j\beta_e + \xi \\ &= -\Gamma_1 + \xi \end{aligned} \tag{2.25}$$

Using (2.24) and (2.25) we will rewrite (2.23) as

$$1 = \frac{-KI_0\beta_e^2(-\beta_e^2 - 2j\beta_e\xi + \xi^2)}{2V_0(2j\beta_e\xi - \xi^2)(\xi^2)} \tag{2.26}$$

Now we will find that, for typical traveling-wave tubes, ξ is much smaller than β_e; hence we will neglect the terms involving $\beta_e\xi$ and ξ^2 in the numerator in comparison with β_e^2 and we will neglect the term ξ^2 in the denominator in comparison with the term involving $\beta_e\xi$. This gives us

$$\xi^3 = -j\beta_e^3\frac{KI_0}{4V_0} \tag{2.27}$$

While (2.27) may seem simple enough, it will later be found very convenient

to rewrite it in terms of other parameters, and we will introduce them now. Let

$$KI_0/4V_0 = C^3 \tag{2.28}$$

C is usually quite small and is typically often around .02. Instead of ξ we will use a quantity or a parameter δ

$$\xi = \beta_e C\delta \tag{2.29}$$

In terms of δ and C, (2.27) becomes

$$\delta = (-j)^{1/3} = (e^{j(2n-1/2)\pi})^{1/3} \tag{2.30}$$

This has three roots which will be called δ_1, δ_2 and δ_3, and these represent three forward waves. They are

$$\begin{aligned} \delta_1 &= e^{-j\pi/6} = \sqrt{3}/2 - j/2 \\ \delta_2 &= e^{-j5\pi/6} = -\sqrt{3}/2 - j/2 \\ \delta_3 &= e^{j\pi/2} = j \end{aligned} \tag{2.31}$$

Figure 2.4 shows the three values of δ. Equation (2.23) was of the fourth degree, and we see that a wave is missing. The missing root was eliminated

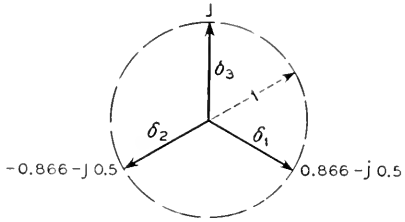

Fig. 2.4—There are three forward waves, with fields which vary with distance as $\exp(-j\beta_e + \beta_e C\delta)z$. The three values of δ for the case discussed, in which the circuit is lossless and the electrons move with the phase velocity of the unperturbed circuit wave, are shown in the figure.

by the approximations made above, which are valid for forward waves only. The other wave is a backward wave and its propagation constant is found to be

$$-\Gamma = j\beta_e \left(1 - \frac{C^3}{4}\right) \tag{2.32}$$

As C is a small quantity, C^3 is even smaller, and indeed the backward wave given by (2.32) is practically the same as the backward wave in the absence of electrons. This is to be expected. In the forward direction, there is a cumulative interaction between wave and the electrons because both are moving

at about the same speed. In the backward direction there is no cumulative action, because the wave and the electrons are moving in the opposite directions.

The variation in the z direction for three forward waves is as

$$\exp -\Gamma z = \exp -j\beta_e z \exp \delta C\beta_e z \tag{2.33}$$

We see that the first wave is an increasing wave which travels a little more slowly than the electrons. The second wave is a decreasing wave which travels a little more slowly than the electrons. The third wave is an unattenuated wave which travels faster than the electrons. It can be shown generally that when a stream of electrons interacts with a wave, the electrons must go faster than the wave in order to give energy to it.

It is interesting to know the ratio of line voltage to line current, or the characteristic impedance, for the three forward waves. This may be obtained from (2.5). We see that the characteristic impedance K_n for the nth wave is given in terms for the propagation constant for the nth wave, Γ_n, by

$$K_n = V/I = jX/\Gamma_n \tag{2.34}$$

In terms of δ_n this becomes

$$K_n = K(1 - \beta_e C\delta_n/\Gamma_1) \tag{2.35}$$

$$K_n = K(1 - jC\delta_n) \tag{2.36}$$

We see that the characteristic impedance for the forward waves differs from the characteristic impedance in the absence of electrons by a small amount proportional to C, and that the characteristic impedance has a small reactive component.

We are particularly interested in the rate at which the increasing wave increases. In a number of wave lengths N, the total increase in db is given by

$$\begin{aligned} &20 \log_{10} \exp [(\sqrt{3}/2)(C)(2\pi N)] \text{ db} \\ &= 47.3\, CN \text{ db} \end{aligned} \tag{2.37}$$

We will see later that the overall gain of the traveling-wave tube with a uniform helix can be expressed in the form

$$G = A + BCN \text{ db} \tag{2.38}$$

Here A is a loss relating voltage associated with the increasing wave to the total applied voltage. This loss may be evaluated and will be evaluated later by a proper examination of the boundary conditions at the input of the tube. It turns out that for the case we have considered

$$G = -9.54 + 47.3\, CN \text{ db} \tag{2.39}$$

In considering circuits for traveling-wave tubes, and in reformulating the theory in more general terms later on, it is valuable to express C in terms of parameters other than the characteristic impedance. Two physically significant parameters are the power flow in the circuit and the electric field associated with it which acts on the electron stream. The ratio of the square of the electric field to the power can be evaluated by physical measurement even when it cannot be calculated. For instance, Cutler[5] did this by allowing the power from a wave guide to flow into a terminated helix, so that the power in the helix was the same as the power in the wave guide. He then compared the field in the helix with the field in the wave guide by probe measurements. The field strength in the wave guide could be calculated in terms of the power flow, and hence Cutler's measurements enabled him to evaluate the field in the helix for a given power flow.

The magnitude of the field is given in terms of the magnitude of the voltage by

$$E = |\Gamma V| \tag{2.40}$$

Here E is taken as the magnitude of the field. The power flow in the circuit is given in terms of the circuit voltage by

$$P = |V|^2/2K \tag{2.41}$$

A quantity which we will use as a circuit parameter is

$$E^2/\beta^2 P = 2K \tag{2.42}$$

Here it has been assumed that we are concerned with low-loss circuits, so that Γ_1^2 can be replaced by the phase constant β^2. Usually, β can be taken as equal to β_e, the electron phase constant, with small error, and in the preceding work this has been assumed to be exactly true in (2.23).

In terms of this new quantity, C is given by

$$C^3 = (2K)(I_0/8V_0) = (E^2/\beta^2 P)(I_0/8V_0) \tag{2.43}$$

If we call V_0/I_0 the beam impedance, C^3 is $\frac{1}{4}$ the circuit impedance divided by the beam impedance. It would have been more sensible to use $E^2/2\beta^2 P$ instead of $E^2/\beta^2 P$. Unfortunately the writer feels stuck with his benighted first choice because of the number of curves and published equations which make use of it.

Besides the circuit impedance, another important circuit parameter is the phase velocity. As the electron velocity is made to deviate from the phase velocity of the circuit, the gain falls off. An analysis to be given later

[5] C. C. Cutler, "Experimental Determination of Helical-Wave Properties," *Proc. IRE*, Vol. 36, pp. 230–233, February 1948.

discloses that the allowable range of velocity Δv is of the order of

$$\Delta v \approx \pm Cu_0 \tag{2.44}$$

Thus, the allowable difference between the phase velocity of the circuit and the velocity of the electrons increases as circuit impedance and beam current are increased and decreases as voltage is increased.

We have illustrated the general method of attack to be used and have introduced some of the important parameters concerned with the circuit and with the overall behavior of the tube. In later chapters, the properties of various circuits suitable for traveling-wave tubes will be discussed in terms of impedance and phase velocity and various cases of interest will be worked out by the methods presented.

CHAPTER III

THE HELIX

Synopsis of Chapter

ANY circuit capable of propagating a slow electromagnetic wave can be used in a traveling-wave tube. The circuit most often used is the helix. The helix is easy to construct. In addition, it is a very good circuit. It has a high impedance and a phase velocity that is almost constant over a wide frequency range.

In this chapter various properties of helices are discussed. An approximate expression for helix properties can be obtained by calculating the properties, not of a helix, but of a helically conducting cylindrical sheet of the same radius and pitch as the helix. An analysis of such a sheet is carried out in Appendix II and the results are discussed in the text.

Parameters which enter into the expressions are the free-space phase constant $\beta_0 = \omega/c$, the axial phase constant $\beta = \omega/v$, where v is the phase velocity of the wave, and the radial phase constant γ. The arguments of various Bessel functions are, for instance, γr and γa, where r is the radial coordinate and a is radius of the helix. The parameters β_0, β and γ are related by

$$\beta^2 = \beta_0^2 + \gamma^2$$

For tightly wound helices in which the phase velocity v is small compared with the velocity of light, γ is very nearly equal to β. For instance, at a velocity corresponding to that of 1,000 volt electrons, γ and β differ by only 0.4%.

Figure 3.1 illustrates two parameters of the helically conducting sheet, the radius a and pitch angle ψ. For an actual helix, a will be taken to mean the mean radius, the radius to the center of the wire.

Figure 3.2 shows a single curve which enables one to obtain γ, and hence β, for any value of the parameter

$$\beta_0 a \cot \psi = \frac{\omega a \cot \psi}{c}.$$

This parameter is proportional to frequency. The curve is an approximate representation of velocity vs. frequency. At high frequencies γ approaches

$\beta_0 \cot \psi$ and β thus approaches $\beta_0/\sin \psi$; this means that the wave travels with the velocity of light around the sheet in the direction of conduction. In the case of an actual helix, the wave travels along the wire with the velocity of light.

The gain parameter C is given by

$$C = (I_0/8V_0)^{1/3}(E^2/\beta^2 P)^{1/3}$$

Values of $(E^2/\beta^2 P)^{1/3}$ on the axis may be obtained through the use of Fig. 3.4, where an impedance parameter $F(\gamma a)$ is plotted vs. γa, and by use of (3.9). For a given helix, $(E^2/\beta^2 P)^{1/3}$ is approximately proportional to $F(\gamma a)$. $F(\gamma a)$ falls as frequency increases. This is partly because at high frequencies and short wavelengths, for which the sign of the field alternates rapidly with distance, the field is strong near the helix but falls off rapidly away from the helix and so the field is weak near the axis. At very high frequencies the field falls off away from the helix approximately as $\exp(-\gamma \Delta r)$, where Δr is distance from the helix, and we remember that γ is very nearly proportional to frequency. $(E^2/\beta^2 P)^{1/3}$ measured at the helix also falls with increasing frequency.

In many cases, a hollow beam of radius r (the dashed lines of Fig. 3.5 refer to such a beam) or a solid beam of radius r (the solid lines of Fig. 3.5 refer to such a beam) is used. For a hollow beam we should evaluate E^2 in $(E^2/\beta^2 P)^{1/3}$ at the beam radius, and for a solid beam we should use the mean square value of E averaged over the beam.

The ordinate in Fig. 3.5 is a factor by which $(E^2/\beta^2 P)^{1/3}$ as obtained from Fig. 3.4 and (3.9) should be multiplied to give $(E^2/\beta^2 P)^{1/3}$ for a hollow or solid beam.

The gain of the increasing wave is proportional to $F(\gamma a)$ times a factor from Fig. 3.5, and times the length of the tube in wavelengths, N. N is very nearly proportional to frequency. Also γ, and hence γa, are nearly proportional to frequency. Thus, $F(\gamma a)$ from Fig. 3.4 times the appropriate factor from Fig. 3.5 times γa gives approximately the gain vs. frequency, (if we assume that the electron speed matches the phase velocity over the frequency range). This product is plotted in Fig. 3.6. We see that for a given helix size the maximum gain occurs at a higher frequency and the bandwidth is broader as r/a, the ratio of the beam radius to the helix radius, is made larger.

It is usually desirable, especially at very short wavelengths, to make the helix as large as possible. If we wish to design the tube so that gain is a maximum at the operating frequency, we will choose a so that the appropriate curve of Fig. 3.6 has its maximum at the value of γa corresponding to the operating frequency. We see that this value of a will be larger the larger is r/a. In an actual helix, the maximum possible value of r/a is less than unity,

since the inside diameter of the helix is less than a by the radius of the wire. Further, focusing difficulties preclude attaining a beam radius equal even to the inside radius of the helix.

Experience indicates that at very short wavelengths (around 6 millimeters, say) it is extremely important to have a well-focused electron beam with as large a value of r/a as is attainable.

A characteristic impedance K_t may be defined in terms of a "transverse" voltage V_t, obtained by integrating the peak radial field from a to ∞, and from the power flow. In Fig. 3.7, $(v/c)\ K_t$ is plotted vs. γa. A "longitudinal" characteristic impedance $K\ell$ is related to K_t (3.13). For slow waves $K\ell$ is nearly equal to K_t. The impedance parameter $E^2/\beta^2 P$ evaluated at the surface of the cylinder is twice $K\ell$. We see that $K\ell$ falls with increasing frequency.

A simplified approach in analysis of the helically conducting sheet is that of "developing" the sheet; that is, slitting it normal to the direction of conduction and flattening it out as in Fig. 3.8. The field equations for such a flattened sheet are then solved. For large values of γa the field is concentrated near the helically conducting sheet, and the fields near the developed sheet are similar to the fields near the cylindrical sheet. Thus the dashed line in Fig. 3.7 is for the developed sheet and the solid line is for a cylindrical sheet.

For the developed sheet, the wave always propagates with the speed of light in the direction of conduction. In a plane normal to the direction of conduction, the field may be specified by a potential satisfying Laplace's equation, as in the case, for instance, of a two-wire or coaxial line. Thus, the fields can be obtained by the solution of an electrostatic problem.

One can develop not only a helically conducting sheet, but an actual helix, giving a series of straight wires, shown in cross-section in Fig. 3.9. In Case I, corresponding to approximately two turns per wavelength, successive wires are $-, +, -, +$ etc.; in case II, corresponding to approximately four turns per wavelength, successive wires are $+, 0, -, 0, +, 0$ etc.

Figures 3.10 and 3.11 illustrate voltages along a developed sheet and a developed helix.

Figure 3.13 shows the ratio, $R^{1/3}$, of $(E^2/\beta^2 P)^{1/3}$ on the axis to that for a developed helically conducting sheet, plotted vs. d/p. We see that, for a large wire diameter d, $(E^2/\beta^2 P)^{1/3}$ may be larger on the axis than for a helically conducting sheet with the same mean radius and hence the same pitch angle and phase velocity. This is merely because the thick wires extend nearer to the axis than does the sheet. The actual helix is really inferior to the sheet.

We see this by noting that the highest value of $(E^2/\beta^2 P)^{1/3}$ for a helically conducting sheet is that at the sheet $(r = a)$. With a finite wire size, the

largest value r can have is the mean helix radius a minus the wire radius. In Fig. 3.14, the ratio of $(E^2/\beta^2P)^{1/3}$ for this largest allowable radius to $(E^2/\beta^2P)^{1/3}$ at the surface of the developed sheet is plotted vs. d/p. We see that, in terms of maximum available field, $(E^2/\beta^2P)^{1/3}$ is no more than 0.83 as high as for the sheet for four turns per wavelength and 0.67 as high as for the sheet for two turns per wavelength. We further see that there is an optimum ratio of wire diameter to pitch; about 0.175 for four turns per wavelength and about 0.125 for two turns per wavelength. Because the maxima are so broad, it is probably better in practice to use larger wire, and in most tubes which have been built, d/p has been around 0.5.

In designing tubes it is perhaps best to do so in terms of field on the axis (Fig. 3.13), the allowable value of r/a and the curves of Fig. 3.6.

Figure 3.15 compares the impedance of the developed helix with that of the developed sheet as given by the straight line of Fig. 3.7.

There are factors other than wire size which can cause the value of E^2/β^2P for an actual helix to be less than the value for the helically conducting sheet. An important cause of impedance reduction is the influence of dielectric supporting members. Even small ceramic or glass supporting rods can cause some reduction in helix impedance. In some tubes the helix is supported inside a glass tube, and this can cause a considerable reduction in helix impedance.

When a field analysis seems too involved, it may be possible to obtain some information by considering the behavior of transmission lines having parameters adjusted to make the phase constant and the characteristic impedance equal to those of the helix. For instance, suppose that the presence of dielectric material results in an actual phase constant β_d as opposed to a computed phase constant β. Equation (3.64) gives an estimate of the consequent reduction of $(E^2/\beta^2P)^{1/3}$ on the axis.

This method is of use in studying the behavior of coupled helices. For instance, concentric helices may be useful in producing radial fields in tubes in which transverse fields predominate in the region of electron flow (see Chapter XIII). A concentric helix structure might be investigated by means of a field analysis, but some interesting properties can be deduced more simply by considering two transmission lines with uniformly distributed self and mutual capacitances and inductances, or susceptance and reactances. The modes of propagation on such lines are affected by coupling in a manner similar to that in which the modes of two resonant circuits are affected by coupling.

If two lines are coupled, their two independent modes of propagation are mixed up to form two modes of propagation in which both lines participate. If the original phase velocities differ greatly, or if the coupling between the lines is weak, the fields and velocity of one of these modes will be almost

like the original fields and velocity of one line, and the fields and velocity of the other mode will be almost like the original fields and velocity of the other line. However, if the coupling is strong enough compared with the original separation of phase velocities, both lines will participate almost equally in each mode. One mode will be a "longitudinal mode" for which the excitations on the two lines are substantially equal, and the other mode will be a "transverse" mode for which the excitations are substantially equal and opposite.

The ratios of the voltages on the lines for the two modes are given by (3.75). Here it is assumed that the series reactances X and shunt susceptances B of the lines are almost equal, differing only enough to make a difference $\Delta\Gamma_0$ in the propagation constants. B_{12} and X_{12} are the mutual susceptance and reactance. We see that to make the voltages on the two lines nearly equal or equal and opposite, B_{12} and X_{12} should have the same sign, so that capacitive and inductive couplings add.

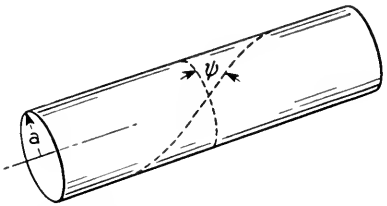

Fig. 3.1—A helically conducting sheet of radius a. The sheet is conducting along helical paths making an angle ψ with a plane normal to the axis.

Increasing the coupling increases the velocity separation between the two modes, and this is desirable. When there is a substantial difference in velocity, operation in the desired mode can be secured by making the electron velocity equal to the phase velocity of the desired mode.

To make the capacitive and inductive couplings add in the case of concentric helices (Fig. 3.17), the helices should be wound in opposite directions.

3.1 The Helically Conducting Sheet

In computing the properties of a helix, the actual helix is usually replaced by a helically conducting cylindrical sheet of the same mean radius. Such a sheet is illustrated in Fig. 3.1. This sheet is perfectly conducting in a helical direction making an angle ψ, the pitch angle, with a plane normal to the axis (the direction of propagation), and is non-conducting in a helical direction normal to this ψ direction, the direction of conduction. Appropriate solutions of Maxwell's equations are chosen inside and outside of the cylindrical sheet. At the sheet, the components of the electric field in the ψ direction are made zero, and those normal to the ψ direction are made equal inside and outside. Since there can be no current in the sheet normal to the ψ direction, the

components of magnetic field in the ψ direction must be the same inside and outside of the sheet. When these boundary conditions are imposed, one can solve for the propagation constant and $E^2/\beta^2 P$ can then be obtained by integrating the Poynting vector.

The helically conducting sheet is treated mathematically in Appendix II. The results of this analysis will be presented here.

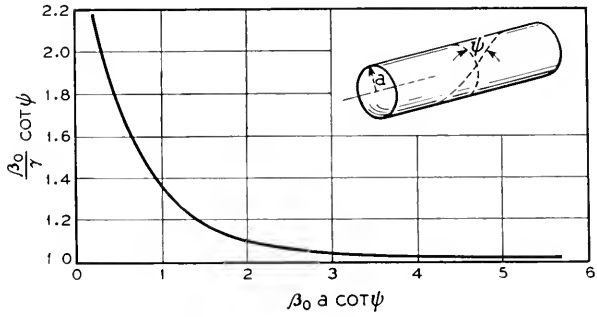

Fig. 3.2—The radial propagation constant is $\gamma^2 = (\beta^2 - \beta_0^2)^{1/2}$. Here (β_0/γ) cot ψ is plotted vs $\beta_0 a$ cot ψ, a quantity proportional to frequency. For slow waves the ordinate is roughly the ratio of the wave velocity to the velocity the wave would have if it traveled along the helically conducting sheet with the speed of light in the direction of conduction.

3.1a *The Phase Velocity*

The results for the helically conducting sheet are expressed in terms of three phase or propagation constants. These are

$$\beta_0 = \omega/c, \qquad \beta = \omega/v \tag{3.1}$$

$$\gamma = \sqrt{\beta^2 - \beta_0^2} \tag{3.2}$$

$$\gamma = \beta\sqrt{1 - (v/c)^2} \tag{3.3}$$

Here c is the velocity of light and v is the phase velocity of the wave. β_0 is the phase constant of a wave traveling with the speed of light, which would vary with distance in the z direction as $\exp(-j\beta_0 z)$. The actual axial phase constant is β, and the fields vary with distance as $\exp(-j\beta z)$.

γ is the radial propagation constant. Various field components vary as modified Bessel functions of argument γr, where r is the radius. Particularly, the longitudinal electric field, which interacts with the electrons, varies as $I_0(\gamma r)$.

For the phase velocities usually used, γ is very nearly equal to β, as may be seen from the following table of accelerating voltages V_0 (to give an electron the velocity v), v/c and γ/β.

V	v/c	γ/β
100	.0198	1.000
1,000	.0625	.998
10,000	.1980	.980

Figure 3.2 gives information concerning the phase velocity of the wave in the form of a plot of $(\beta_0/\gamma)\cot\psi$ as a function of $\beta_0 a \cot\psi$.

The ratio of the phase velocity v to the velocity of light c may be expressed

$$v/c = \beta_0/\beta = (\gamma/\beta)(\beta_0/\gamma)\cot\psi\tan\psi \tag{3.4}$$

$$v/c = (\gamma/\beta)\tan\psi\,[(\beta_0/\gamma)\cot\psi\,]$$

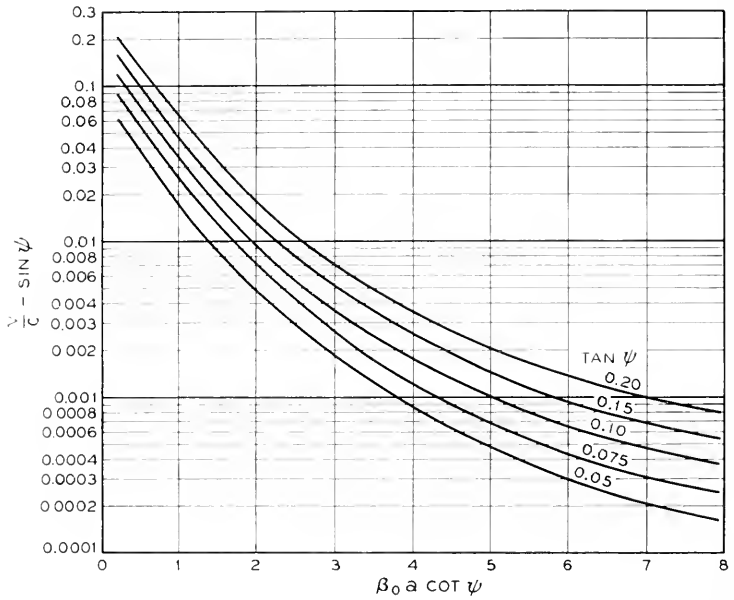

Fig. 3.3 – From these curves one can obtain v/c, the ratio of the phase velocity of the wave to the velocity of light, for various values of tan ψ and $\beta_0 a$ cot ψ.

From Fig. 3.2 we see that, for large values of $\beta_0 a$ cot ψ, (β_0/γ) cot ψ approaches unity. For slow waves γ/β approaches unity. Under these circumstances, very nearly

$$v/c = \tan\psi \tag{3.5}$$

If the wave traveled in the direction of conduction with the speed of light we would have

$$v/c = \sin\psi$$

This is essentially the same as (3.5) for small pitch angles ψ. Thus, for large values of the abscissa in Fig. 3.2, the phase velocity is just about that corresponding to propagation along the sheet in the direction of conduction with the speed of light and hence in the axial direction at a much reduced speed. For helices of smaller radius compared with the wavelength, the speed is greater.

The bandwidth of a traveling-wave tube is in part determined by the range over which the electrons keep in step with the wave. The abscissa of Fig. 3.2 is proportional to frequency, but the ordinate is not strictly proportional to phase velocity. Hence, it seems desirable to have a plot which does show velocity directly. To obtain this we can assign various values to cot ψ.

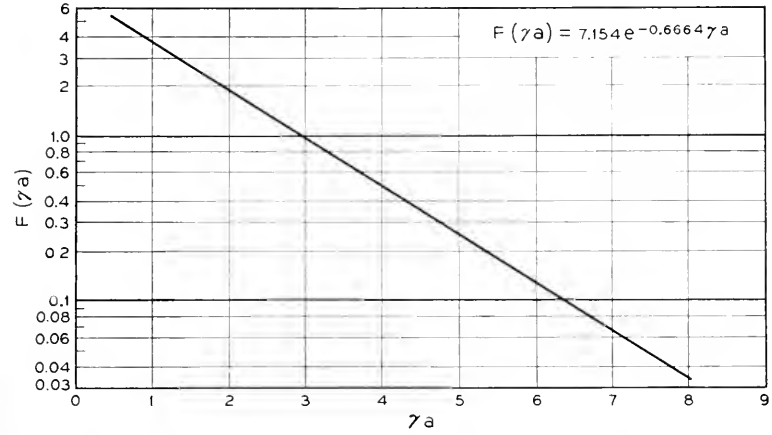

Fig. 3.4—A curve giving the impedance function $F(\gamma a)$ vs. γa. On the axis, $(E^2/\beta^2 P)^{1/3} = (\beta/\beta_0)^{1/3}(\gamma/\beta)^{1/3}F(\gamma a)$.

The ordinate (β_0/γ) cot ψ then gives us γ/β_0 and from (3.2) we see that

$$v/c = \beta_0/\beta = (1 + (\gamma/\beta_0)^2)^{-1/2} \tag{3.6}$$

We have seen that, for large values of $\beta_0 a$ cot ψ, (β_0/γ) cot ψ approaches unity, and v/c approaches a value

$$v/c = (1 + \cot^2 \psi)^{-1/2} = \sin \psi \tag{3.7}$$

To emphasize the change in velocity with frequency it seems best to plot the difference between the actual velocity ratio and this asymptotic velocity ratio on a semi-log scale. Accordingly, Fig. 3.3 shows $(v/c) - \sin \psi$ vs. $\beta_0 a$ cot γ for tan ψ = .05, .075, .1, .15, .2.

For large values of the abscissa the velocities are those corresponding to

about 640 volts ($\tan \psi = .05$), 1,400 volts (.075), 2,500 volts (.1), 5,600 volts (.15), 9,800 volts (.2).

3.1b *The Impedance Parameter* ($E^2/\beta^2 P$)

Figure 3.4 shows a plot of a quantity $F(\gamma a)$ vs. γa. This quantity is computed from a very complicated expression (Appendix II), but it is accurately given over the range shown by the empirical relation

$$F(\gamma a) = 7.154\, e^{-.6664\gamma a} \tag{3.8}$$

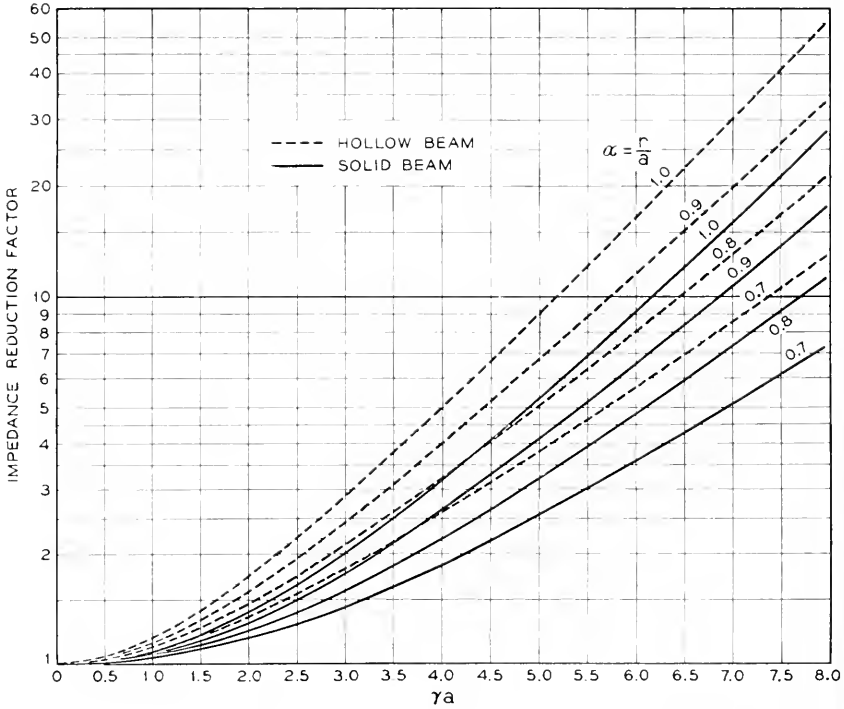

Fig. 3.5—Factors by which $(E^2\beta/^2P)^{1/3}$ on the axis should be multiplied to give the correct value for hollow and solid beams of radius r.

For the field on the axis of the helix,

$$(E^2/\beta^2 P)^{1/3} = (\beta/\beta_0)^{1/3}(\gamma/\beta)^{4/3} F(\gamma a) \tag{3.9}$$

We should remember that $\beta/\beta_0 = c/v$ and that γ/β is nearly unity for velocities small compared with the velocity of light.

In the expression for the gain parameter C, the square of the field E is multiplied by the current I_0 (2.28). If we were to assume that two electron

streams of different currents, I_1 and I_2, were coupled to the circuit through transformers, so as to be acted on by fields E_1 and E_2, but that the streams did not interact directly with one another, we would find the effective value of C^3 to be given by

$$C^3 = (E_1^2/\beta^2 P)(I_1/8V_0) + (E_2^2/\beta^2 P)(I_2/8V_0)$$

Thus, if we neglect the direct interaction of electron streams through fields due to local space charge, we can obtain an effective value of C^3 by integrating $E^2 dI_0$ over the beam. If we assume a constant current density, we can merely use the mean square value of E over the area occupied by electron flow.

The axial component of electric field at a distance r from the axis is $I_0(\gamma r)$ times the field on the axis. Hence, if we used a tubular beam of radius r, we should multiply $(E^2/\beta^2 P)^{1/3}$ as obtained from Fig. 3.4 by $[I_0(\gamma r)]^{2/3}$. The quantity $[I_0(\gamma r)]^{2/3}$ is plotted vs. γa for several values of r/a as the dashed lines in Fig. 3.5.

Suppose the current density is uniform out to a radius r and zero beyond this radius. The average value of E^2 is greater than the value on the axis by a factor $[I_0^2(\gamma r) - I_1^2(\gamma r)]$ and $(E^2/\beta^2 P)^{1/3}$ from Fig. 3.4 should in this case be multiplied by this factor to the $\frac{1}{3}$ power. The appropriate factor is plotted vs. γa as the solid lines of Fig. 3.5.

We note from (2.39) that the gain contains a term proportional to CN, where N is the number of wavelengths. For slow waves and usual values of γa, very nearly, N will be proportional to the frequency and hence to γ, while C is proportional to $(E^2/\beta^2 P)^{1/3}$. We can obtain $(E^2/\beta^2 P)^{1/3}$ from Figs. 3.4 and 3.5. The gain of the increasing wave as a function of frequency will thus be very nearly proportional to this value of $(E^2/\beta^2 P)^{1/3}$ times γ, or, times γa if we prefer.

In Fig. 3.6, $\gamma a F(\gamma a)$ is plotted vs. γa for hollow beams of radius r for various values of r/a (dashed lines) and for uniform density beams of radius r for various values of r/a (solid lines). If we assume that the electron speed is adjusted to equal the phase velocity of the wave, we can take the ordinate as proportional to gain and the abscissa as proportional to frequency.

We see that the larger is r/a, the larger is the value of γa for maximum gain. For one typical 7.5 cm wavelength traveling-wave tube, γa was about 2.8. For this tube, the ratio of the inside radius of the helix to the mean radius of the helix was 0.87. We see from Fig. 3.6 that, if a solid beam just filled this helix, the maximum gain should occur at about the operating wavelength. As a matter of fact, the beam was somewhat smaller than the inside diameter of the helix, and there was an observed increase of gain with an increase in wavelength (a higher gain at a lower frequency). In a particular

tube for 0.625 cm wavelength, it was felt desirable to use a relatively large helix diameter. Accordingly, a value of γa of 6.7 was chosen. We see that, unless r/a is 0.9 or larger, this must result in an appreciable increase in gain at some frequency lower than operating frequency. It was only by use of great care in focusing the beam that gain was attained at 0.625 cm wavelength, and there was a tendency toward oscillation, presumably at longer wavelengths. This discussion of course neglects the effect of transmission

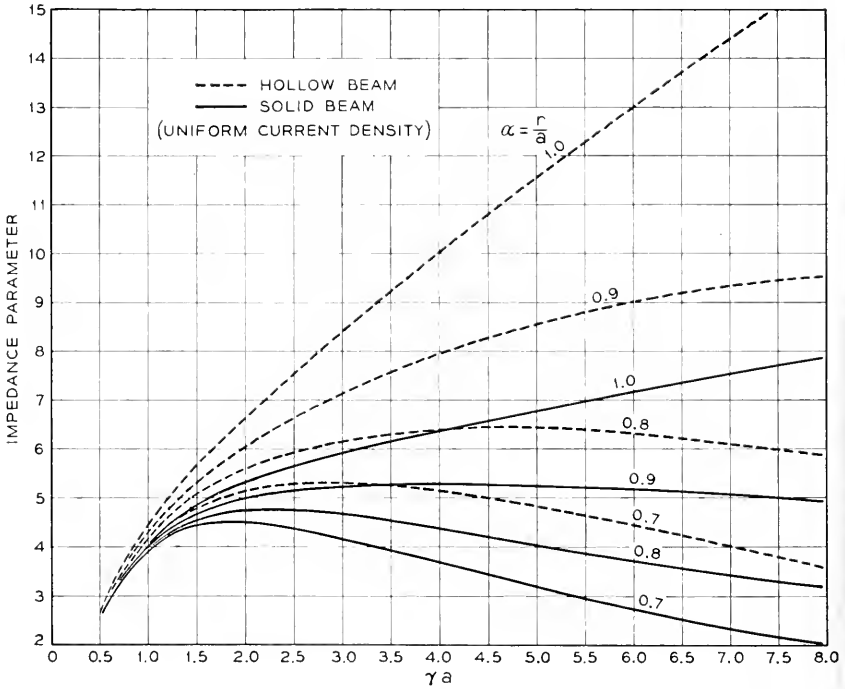

Fig. 3.6—The ordinate is $\gamma aF(\gamma a)$ times the parameters from Fig. 3.5. For a fixed current and voltage it is nearly proportional to gain per unit length, and hence the curves give roughly the variation of gain with frequency.

loss or gain. Usually the loss decreases when the frequency is decreased, and this favors oscillation at low frequencies.

3.1c *Impedance of the Helix*

No impedance which can be assigned to the helically conducting sheet can give full information for matching a helix to a waveguide or transmission line. As in the case of transducers between a coaxial line and a waveguide or between waveguides of different cross-section, the impedance is important,

but discontinuity effects are also important. However, a suitably defined helix impedance is of some interest.

Figure 3.7 presents the impedance as defined on a voltage-power basis. The peak "transverse" voltage V_t is obtained by integrating the radial electric field from the radius a of the helically conducting sheet to ∞. The "transverse" characteristic impedance K_t is defined by the relation

$$P = (\tfrac{1}{2})(V_t^2/K_t)$$

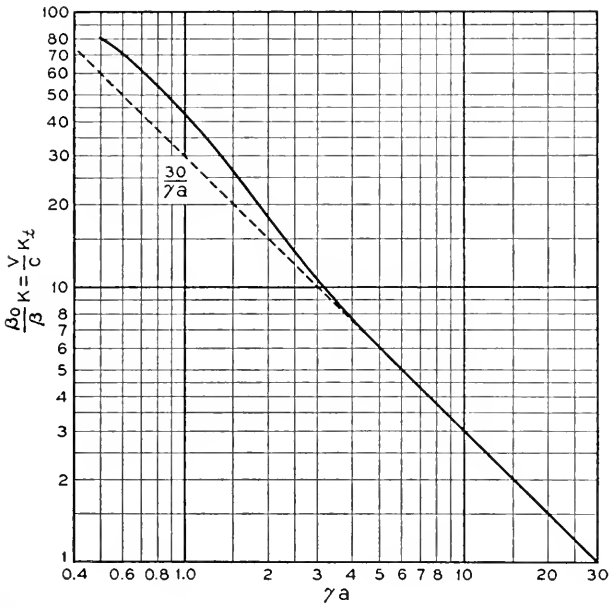

Fig. 3.7—Curves giving the variation of transverse impedance, K_t, with γa.

The impedance is found to be given by

$$\left(\frac{\beta}{\gamma}\right)^2 \left(\frac{\beta_0}{\beta}\right) K_t = \frac{120 I_0^2}{(\gamma a)^2} \left[\left(1 + \frac{I_0 K_1}{I_1 K_0}\right)(I_1^2 - I_0 I_2) + \left(\frac{I_0}{K_0}\right)^2 \left(1 + \frac{I_1 K_0}{I_0 K_1}\right)(K_0 K_2 - K_1^2) \right]^{-1} \tag{3.10}$$

The I's and K's are modified Bessel functions of argument γa.

The dashed line on Fig. 3.7 is a plot of $30/\gamma a$ vs. γa. It may be seen that, for large values of γa, very nearly

$$K_t = (\beta/\beta_0)(\gamma/\beta)^2(30/\gamma a) \tag{3.11}$$

and in the whole range shown the impedance differs from this value by a factor less than 1.5.

We might have defined a "longitudinal" voltage $V\ell$ as half of the integral of the longitudinal component of electric field at the surface of the helically conducting sheet for a half wavelength (between successive points of zero field). We find that

$$V\ell = \sqrt{1 - (v/c)^2}\, V_t = (\gamma/\beta)V_t \tag{3.12}$$

and, accordingly, the "longitudinal impedance" $K\ell$ will be

$$K\ell = [1 - (v/c)^2]K_t = (\gamma/\beta)^2 K_t \tag{3.13}$$

Our impedance parameter, $E^2/\beta^2 P$, is just twice this "longitudinal impedance."

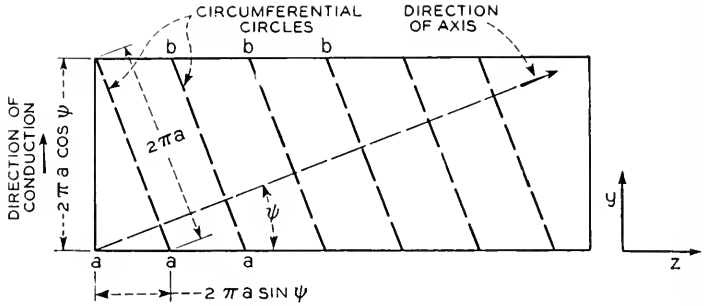

Fig. 3.8—A "developed" helically conducting sheet. The sheet has been slit along a line normal to the direction of conduction and flattened out.

The transverse voltage V_t is greater than the longitudinal voltage $V\ell$ because of the circumferential magnetic flux outside of the helix. For slow waves $V\ell$ is nearly equal to V_t and the fields are nearly curl-free solutions of Laplace's equation. In this case the circumferential magnetic flux is small compared with the longitudinal flux inside of the helix.

For the circuit of Fig. 2.3 the transverse and longitudinal voltages are equal, and it is interesting to note that this is approximately true for slow waves on a helix. For very fast waves, the longitudinal voltage becomes small compared with the transverse voltage.

For a typical 4,000-megacycle tube, for which $\gamma a = 2.8$, Fig. 5 indicates a value of K_t of about 150 ohms.

3.2 The Developed Helix

For large helices, i.e., for large values of γa, the fields fall off very rapidly away from the wire. Under these circumstances we can obtain quite accurate results by slitting the helically conducting sheet along a spiral line normal

to the direction of conduction and flattening it out. This gives us the plane conducting sheet shown in Fig. 3.8. The indicated coordinates are z to the right and y upward: x is positive into the paper. The fields about the developed sheet approximate those about the helically conducting sheet for distances always small compared with the original radius of curvature.

The straight dashed line shown on the helix impedance curve of Fig. 3.7 can be obtained as a solution for the "developed helix." We see that it is within 10% of the true curve for values of γa greater than 2.8. We might note that a 10% error in impedance means only a $3\frac{1}{3}\%$ error in the gain parameter C.

In solving for the fields around the sheet, the developed surface can be extended indefinitely in the plus and minus y directions. In order that the fields may match when the sheet is rolled up, they must be the same at $y = 0, z = 2\pi a \sin\psi$ and $y = 2\pi a \cos\psi, z = 0$. The appropriate solutions are plane electromagnetic waves traveling in the y direction with the speed of light.

For positive values of x, the appropriate electric and magnetic fields are

$$\begin{aligned} E_x &= E_0 e^{-\gamma x} e^{-j\gamma z} e^{-j\beta_0 y} \\ E_z &= jE_0 e^{-\gamma x} e^{-j\gamma z} e^{-j\beta_0 y} \\ E_y &= 0 \end{aligned} \tag{3.14}$$

We should note that the x and z components of the field can be obtained as gradients of a function

$$\Phi = -(E_0/\gamma) e^{-\gamma x} e^{-j\gamma z} e^{-j\beta_0 y} \tag{3.15}$$

where

$$\begin{aligned} E_x &= -\partial\Phi/\partial z \\ E_z &= -\partial\Phi/\partial y \end{aligned} \tag{3.16}$$

$$\partial^2\Phi/\partial x^2 + \partial^2\Phi/\partial z^2 = 0 \tag{3.17}$$

Thus, in the xz plane, Φ satisfies Laplace's equation.

The magnetic field is given by the curl[6] of the electric field times $j/\omega\mu$. Its components are:

$$\begin{aligned} H_x &= \frac{-j}{\mu c} E_0 e^{-\gamma x} e^{-j\gamma z} e^{-j\beta_0 y} \\ H_z &= \frac{-1}{\mu c} E_0 e^{-\gamma x} e^{-j\gamma z} e^{-j\beta_0 y} \\ H_y &= 0 \end{aligned} \tag{3.18}$$

[6] Maxwell's equations are given in Appendix I.

The fields in the $-x$ direction may be obtained by substituting $\exp(\gamma x)$ for $\exp(-\gamma x)$.

If the sheet is to roll up properly, the points a on the bottom coinciding with the points b on the top, we have

$$2\pi\gamma a \sin\psi - 2\pi\beta_0 a \cos\psi = 2n\pi \tag{3.19}$$

where n is an integer.

The solution corresponding most nearly to the wave on a singly-wound helix is that for $n = 0$. The others lead to a variation of field by n cycles along a circumferential line. These can be combined with the $n = 0$ solution to give a solution for a developed helix of thin tape, for instance. Or, appropriate combinations of them can represent modes of helices wound of several parallel wires. For instance, we can imagine winding a balanced transmission line up helically. One of the modes of propagation will be that in which the current in one wire is 180° out of phase with the current in the other. This can be approximated by a combination of the $n = +1$ and $n = -1$ solutions. This mode should not be confused with a fast wave, a perturbation of a transverse electromagnetic wave, which can exist around an unshielded helix.

Usually, we are interested in the slow wave on a singly-wound helix, and in this case we take $n = 0$ in (3.19), giving

$$\begin{aligned} \gamma \sin\psi - \beta_0 \cos\psi &= 0 \\ \tan\psi &= \beta_0/\gamma \end{aligned} \tag{3.20}$$

$$\sin\psi = \frac{\beta_0}{(\gamma^2 + \beta_0^2)^{1/2}} \tag{3.21}$$

$$\cos\psi = \frac{\gamma}{(\gamma^2 + \beta_0^2)^{1/2}} \tag{3.22}$$

Let us evaluate the propagation constant in the axial direction. From Fig. 3.8 we see that, in advancing unit distance in the axial direction, we proceed a distance $\cos\psi$ in the z direction and $\sin\psi$ in the y direction. Hence, the phase constant β in the axial direction must be

$$\beta = \beta_0 \sin\psi + \gamma \cos\psi \tag{3.23}$$

Using (3.18) and (3.19), we obtain

$$\beta = (\beta_0^2 + \gamma^2)^{1/2} \tag{3.24}$$

$$\gamma = (\beta^2 - \beta_0^2)^{1/2} \tag{3.25}$$

These are just relations (3.2, 3.3).

The power flow along the axis is that crossing a circumferential circle, represented by lines *a-b* in Fig. 3.8. As the power flows in the y direction, this is the power associated with a distance $2\pi a \sin \psi$ in z direction. Also, the power flow in the $+x$ region will be equal to the power flow in the $-x$ region. Hence, the power flow in the helix will be twice that in the region $x = 0$ to $x = +\infty$, $z = 0$ to $z = 2\pi a \sin \psi$.

$$P = 2 \int_{z=0}^{2\pi a \sin \psi} \int_{x=0}^{\infty} (\tfrac{1}{2})(E_z H_x^* - E_x H_z^*)\, dx\, dz \tag{3.26}$$

This is easily integrated to give

$$P = \frac{2\pi a \sin \psi E_0^2}{\gamma \mu c} \tag{3.27}$$

The magnitude E of the axial component of field is

$$E = E_0 \cos \psi \tag{3.28}$$

Using (3.21), (3.22), (3.24) and (3.28) in connection with (3.27) we obtain

$$(E^2/\beta^2 P) = (\gamma/\beta)^4(\beta/\beta_0)(\mu c/2\pi\gamma a) \tag{3.29}$$

We have

$$\mu c = \mu/\sqrt{\mu\epsilon} = \sqrt{\mu/\epsilon} = 377 \text{ ohms}$$

Thus

$$E^2/\beta^2 P = (\gamma/\beta)^4(\beta/\beta_0)(60/\gamma a) \tag{3.30}$$

The longitudinal impedance is half this, and the transverse impedance is $(\beta/\gamma)^2$ times the longitudinal impedance.

3.3 Effect of Wire Size

An actual helix of round wire, as used in traveling-wave tubes, will of course differ somewhat in properties from the helically conducting sheet for which the foregoing material applies.

One might expect a small difference if there were many turns per wavelength, but actual tubes often have only a few turns per wavelength. For instance, a typical 4,000 mc tube has about 4.8 turns per wavelength, while a tube designed for 6 mm operation has 2.4 turns per wavelength.

If the wire is made very small there will be much electric and magnetic energy very close to the wire, which is not associated with the desired field component (that which varies as $\exp(-j\beta z)$ in the z direction). If the wire is very large the internal diameter of the helix becomes considerably less than the mean diameter, and the space available for electron flow is reduced. As the field for the helically conducting sheet is greatest at the sheet, this

means that the maximum available field is reduced. Too, the impedance will depend on wire size.

It thus seems desirable to compare in some manner an actual helix and the helically conducting sheet. It would be very difficult to solve the problem of an actual helix. However, we can make an approximate comparison by a method suggested by R. S. Julian.

In doing this we will develop the helix of wires just as the helically con-

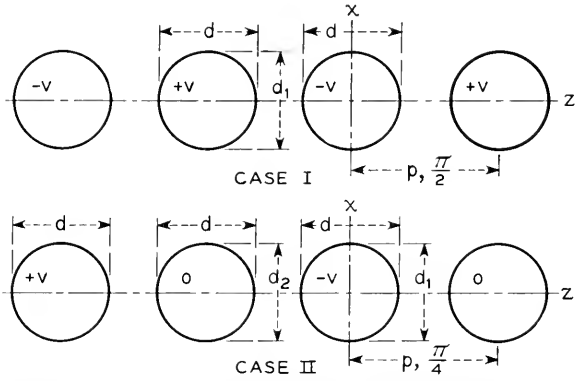

Fig. 3.9—The wires of a developed helix with about two turns per wavelength (case I) and about four turns per wavelength (case II). In the analysis used, the wires are not quite round.

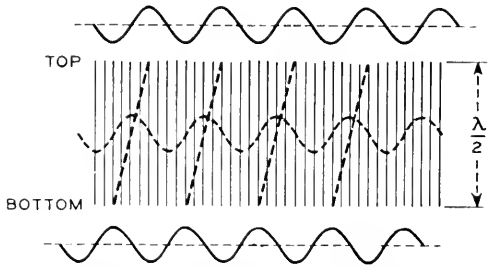

Fig. 3.10—Voltages on a developed helically conducting sheet for two turns per wavelength.

ducting sheet was developed, by slitting it along a helical line normal to the wires. We will then consider two special cases, one in which the wires of the developed helix are one half wavelength long and the other in which the wires are one quarter wavelength long.

The waves propagated on the developed helix are transverse electromagnetic waves propagated in the direction of the wires, and the electric fields normal to the direction of propagation can be obtained from a solution of Laplace's equation in two dimensions (as in (3.15)–(3.17)).

It is easy to make up two-dimensional solutions of Laplace's equation with equipotentials or conductors of approximately circular form, as shown in Fig. 3.9. In case I, the conductors are alternately at potentials $-V, +V, -V$, etc.; and in case II, the potentials are $-V$, 0, $+V$, 0, $-V$, 0, $+V$, etc. Far away in the x direction from such a series of conductors, the field will vary sinusoidally in the z direction and will vary in the same manner with x as in the developed helically conducting sheet. Hence, we can make the distant fields of the conductors of cases I and II of Fig. 3.9 equal to the distant fields of developed helically conducting sheets, and compare the $E^2/\beta^2 P$ and the impedance for the different systems. Case I would correspond to a helix of approximately two turns per wavelength and case II to four turns per wavelength.

3.3a *Two Turns per Wavelength*

Figure 3.10 is intended to illustrate the developed helically conducting sheet. The vertical lines indicate the direction of conduction. The dashed slanting lines are intersections of the original surface with planes normal to the axis. That is, on the original cylindrical surface they were circles about the surface, and they connect positions along the top and bottom which should be brought together in rolling up the flattened surface to reconstitute the helically conducting sheet.

Waves propagate on the developed sheet of Fig. 3.10 vertically with the speed of light. The vertical dimension of the sheet is in this case taken as $\lambda/2$, where λ is the free-space wavelength.[7] The sine waves above and below Fig. 3.10 indicate voltages at the top and the bottom and are, of course, 180° out of phase. As is necessary, the voltages at the ends of the dashed slanting lines, (really, the voltages at the same point before the sheet was slit) are equal.

A wave sinusoidal at the bottom of the sheet, zero half way up and 180° out of phase with the bottom at the top would constitute along any horizontal line a standing wave,not a traveling wave. Actually, this is only one component of the field. The other is a wave 90° out of phase in both the horizontal and vertical directions. Its maximum voltage is half-way up, and it is indicated by the dotted sine wave in Fig. 3.10. The voltage of this component is zero at top and bottom. It may be seen that these two components propagating upward together constitute a wave traveling to the right. The two components are orthogonal spatially, and the total power is twice the power of either component taken separately.

Figure 3.11 indicates an array of wires obtained by developing an actual

[7] Section 3.3a is referred to as "two turns per wavelength." This is not quite accurate; it is in error by the difference between the lengths of the vertical and the slanting lines in Fig. 3.10.

helix which has been slit along a helical line normal to the wire of which the helix is wound. The dashed slanting lines again connect points which were the same point before the helix was slit and developed. Again we assume a height of a half wavelength. Thus, if the polarities are maximum $+, -, +, -$ etc. as shown at the bottom, they will be maximum $-, +, -, +, -, +$ etc. as shown at the top, and zero half-way up. In this case the field is a standing wave along any horizontal line, and no other component can be introduced to make it a traveling wave. Half of the field strength can be regarded as constituting a component traveling to the right and half as a component traveling to the left.

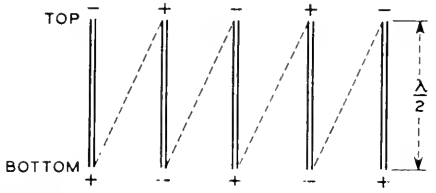

Fig. 3.11—Voltages on a developed helix for two turns per wavelength.

The equipotentials used to represent the field about the wires of Fig. 3.9, Case I and Fig. 3.10 belong to the field

$$V + j\psi = \ln \tan (z + jx) \tag{3.31}$$

Here V is potential and ψ is a stream function. There are negative equipotentials about $z = x = 0$ and positive equipotentials about $x = 0$, $z = \pm\pi/2$. For an equipotential coinciding with the surface of a wire of z-diameter, $2z_{\text{wire}}$, d/p is thus

$$d/p = \frac{z_{\text{wire}}}{\pi/4} \tag{3.32}$$

at $x = 0, z < \pi/4$

$$V = \ln \tan z \tag{3.33}$$

at $z = 0$

$$V = \ln \tanh x \tag{3.34}$$

Hence, for an equipotential on the wire with an z-diameter $2z$, the x-diameter $2x$ can be obtained from (3.33) and (3.34) as

$$2x = 2 \tanh^{-1} \tan z \tag{3.35}$$

Of course, the ratio of the x-diameter d_1 to the pitch is given by

$$d_1/p = \frac{x}{\pi/4} \tag{3.36}$$

where x is obtained from (3.35).

In Fig. 3.12, d_1/d is plotted vs. d/p by means of (3.35) and (3.36). This shows that for wire diameters up to $d/p = .5$ (open space equal to wire diameter) the equipotentials representing the wire are very nearly round.

The total electric flux from each wire is $2\pi\epsilon$ and the potential of a wire of z-diameter $2z$ is $V = -\ln \tan z$. Hence, the stored energy W_1 per unit length per wire, half the product of the charge and the voltage, is

$$W_1 = -\pi\epsilon \ln \tan z \tag{3.37}$$

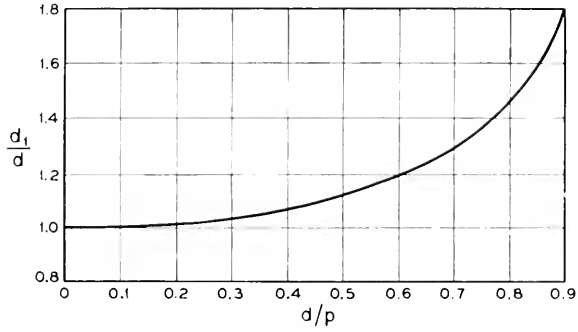

Fig. 3.12—Ratio of the two diameters of the wire of a helix for two turns per wavelength (see Fig. 3.9) vs. the ratio of one of the diameters to the pitch.

The total distant field and the useful field component are given by expanding (3.31) in Fourier series and taking the fundamental component, giving

$$V = -2 \cos 2z e^{\mp 2x} \tag{3.38}$$

The $-$ sign applies for $x > 0$ and the $+$ sign for $x < 0$. Half of this can be regarded as belonging to a field moving to the right and half to a field moving to the left.

For a field equal to half that specified by (3.38), which might be part of the field of a developed helically conducting sheet, the stored energy W_2 per unit depth can be obtained by integrating $(E_z^2 + E_x^2)\ \epsilon/2$ from $x = -\infty$ to $x = +\infty$ and from $z = -\pi/4$ to $+\pi/4$, and it turns out to be

$$W_2 = \tfrac{1}{2}\ \pi\epsilon \tag{3.39}$$

If we add another field component similar to half of (3.38), but in quadrature with respect to z and t, we will have the traveling wave of a helically conducting sheet with the same distant traveling field component as given by (3.31). Hence, the ratio R of the stored energy for the developed sheet to the stored energy for the developed helix is

$$R = 2W_2/W_1 = -\frac{1}{\ln \tan z} \tag{3.40}$$

R is the ratio of the stored energies, and hence of the power flows (since the waves both propagate with the speed of light) of a developed helically conducting sheet and a developed helix with the same distant traveling fundamental field components. Hence, at a given distance $(E^2/\beta^2 P)^{1/3}$ for the helix is $R^{1/3}$ times as great as for the helically conducting sheet. In Fig. 3.13, $R^{1/3}$ is plotted vs. d/p.

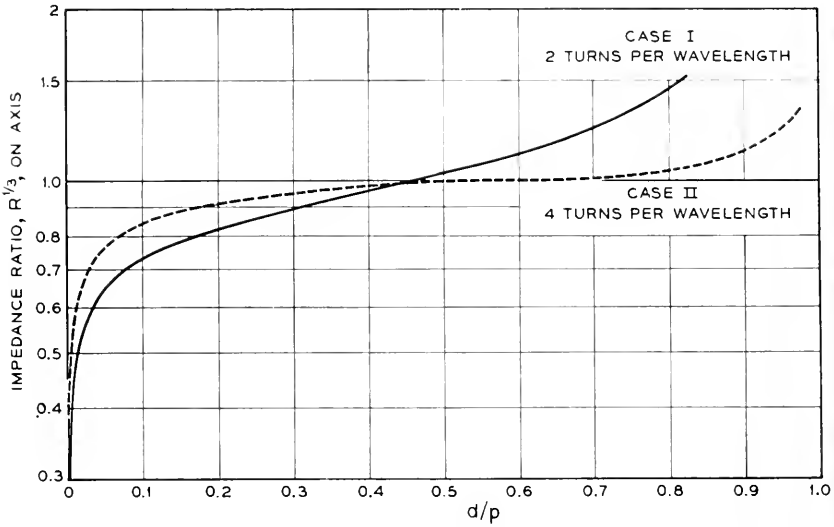

Fig. 3.13—Ratio $R^{1/3}$ of $(E^2/\beta^2 P)^{1/3}$ for a helix to the value for a helically conducting sheet for the distant field.

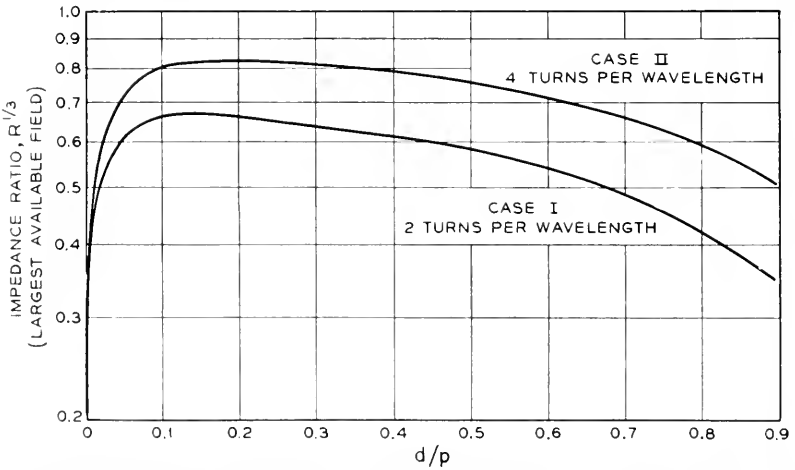

Fig. 3.14—Ratio $R^{1/3}$ of $(E^2/\beta^2 P)^{1/3}$ for a helix to the value for a helically conducting sheet, field at the inside diameter of the helix or sheet.

The maximum available field for the developed helically conducting sheet (equation (3.38)) is that for $x = 0$. The maximum available field for the developed helix (equation (3.31)) is that for an electron grazing the helix inner or outer diameter, that is, an electron at a value of x given by (3.35). The fundamental sinusoidal component of the field varies as $\exp(-2x)$ for both the sheet and the helix, and hence there is a loss in E^2 by a factor $\exp(-4x)$ because of this. We wish to make a comparison on the basis of E^2 and power or energy. Hence, on basis of maximum available field squared we would obtain from (3.40)

$$R = -\frac{1}{\ln \tan z} e^{-4x} \tag{3.41}$$

where x is obtained from (3.35). Figure 3.14 was obtained from (3.32), (3.35) and (3.41).

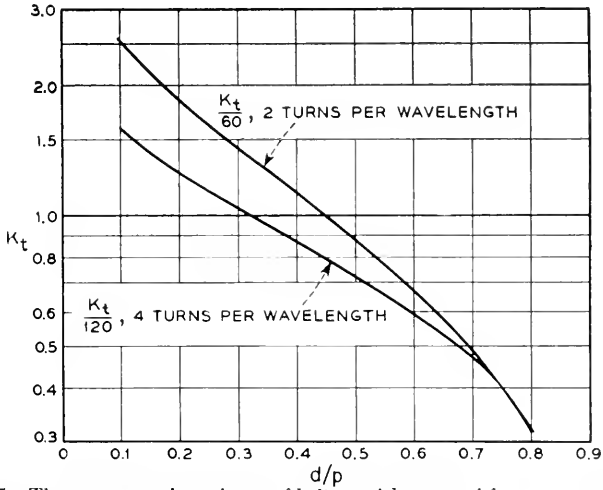

Fig. 3.15—The transverse impedance of helices with two and four turns per wavelength vs. the ratio of wire diameter to pitch.

In a transmission line the characteristic impedance is given by

$$K = \sqrt{\frac{L}{C}} \tag{3.42}$$

Here L and C are the inductance and capacitance per unit length. This impedance should be identified with the transverse impedance of the helix. We also have for the velocity of propagation, which will be the velocity of light, c,

$$c = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{\mu\epsilon}} \tag{3.43}$$

From (3.42) and (3.43) we obtain

$$\begin{aligned} K_t &= \sqrt{\mu\epsilon}/C = \sqrt{\mu/\epsilon}(\epsilon/C) \\ &= 377\ \epsilon/C \end{aligned} \tag{3.44}$$

Now C is the charge Q divided by the voltage V. Hence

$$K_t = 377\ \epsilon V/Q \tag{3.45}$$

In this case we have

$$\begin{aligned} K_t &= \frac{337\epsilon \ln \tan z}{2\pi\epsilon} \\ K_t &= -60 \ln \tan_z \end{aligned} \tag{3.46}$$

To obtain the impedance of the corresponding helically conducting sheet we assume, following (3.30)

$$K_t = (\gamma/\beta)\ (\gamma/\beta_0)\ (30/\gamma a) \tag{3.47}$$

and assuming a slow wave, let $\gamma = \beta$, so that

$$K_t = 30/\beta_0 a \tag{3.48}$$

If we are to have n turns per wavelength, and the speed of light in the direction of conduction, then we must have

$$\beta_0 a = 1/n \tag{3.49}$$

whence

$$K_t = 30n \tag{3.50}$$

For $n = 2$ (two turns per wavelength), $K = 60$. In Fig. 3.15, the characteristic impedance K_t as obtained from (3.46) divided by 60 (from (3.50)) is plotted vs. d/p.

3.3b *Four Turns per Wavelength*

In this case there are enough wires so that we can add a quadrature component as in Fig. 3.10 and thus produce a traveling wave rather than a standing wave. Thus, we can make a more direct comparison between the developed sheet and the developed helix.

For the developed helix we have

$$V + j\psi = \ln \tan (z + jx) + \frac{A}{\cos 2\ (z + jx)} \tag{3.15}$$

If we transform this to new coordinates z_1, x_1 about an origin at $z = 0$, $x = \pi/4$ we obtain

$$V + j\psi = \ln\left(\frac{1 + \tan(z_1 + jx_1)}{1 - \tan(z_1 + jx_1)}\right) - \left(\frac{A}{\sin 2(z_1 + jx_1)}\right) \tag{3.52}$$

We can now adjust A to give a zero equipotential of diameter $2z_1$ about $x = x_1 = 0$, $z_1 = 0$ ($z = \pi/4$) by letting

$$A = (\sin 2z_1) \ln\left(\frac{1 + \tan z_1}{1 - \tan z_1}\right) \tag{3.53}$$

If A is so chosen, there will be roughly circular equipotentials of z-diameter $2z_1$ about $z = \pm \pi/4$, etc. There will also be roughly circular equipotentials of the same z-diameter about $z = 0$, $\pm\pi/2$, etc., of potential $\pm V$. That about $z = 0$ has a potential

$$V = \ln\left(\frac{1 + \tan z_1}{1 - \tan z_1}\right)\frac{A}{\cos 2z_1} \tag{3.54}$$

where A is taken from (3.53).

The distance between centers of equipotentials is $p = \pi/4$, so that the ratio of z-diameter of the equipotentials to pitch is

$$d/p = 2z_1/(\pi/4) = z_1/(\pi/8) \tag{3.55}$$

The x-diameter of the equipotential about $z = 0$ (and of those about $z = \pm\frac{\pi}{2}$ etc.) can be obtained as $2x$ by letting V have the value given by (3.54) and setting $z = 0$ in (3.51), giving

$$V = \ln \tanh x + \frac{A}{\cosh 2x} \tag{3.56}$$

The ratio of this x-diameter to the pitch, d_1/p, is

$$d_1/p = y/(\pi/8), \tag{3.57}$$

x is obtained from (3.56).

To obtain the x-diameter of the 0 potential electrodes we take the derivative (3.52) with respect to z_1, giving the gradient in the z direction

$$\frac{\partial V}{\partial z_1} + j\frac{\partial \psi}{\partial z_1} = \frac{\sec^2(z_1 + jx_1)}{1 + \tan(z_1 + jx_1)} + \frac{\sec^2(z_1 + jx_1)}{1 - \tan(z_1 + jx_1)} - \frac{2A\cos 2(z_1 + jx_1)}{\sin 2(z_1 + jx_1)} \tag{3.58}$$

We then let $z_1 = 0$ and find the value of x_1 for which $\partial V/\partial z_1 = 0$. When $z_1 = 0$, (3.58) becomes

$$A = \sinh 2x_1 \tanh 2x_1 \frac{(1 - \tanh^2 x_1)}{(1 + \tanh^2 x_1)} \tag{3.59}$$

As A is given by (3.53), we can obtain x, from (3.57), and the ratio of the x-diameter d_2 to the pitch is

$$d_2/p = x_1/(\pi/8) \tag{3.60}$$

Figure 3.16 shows d_1/d and d_2/d vs. d/p.

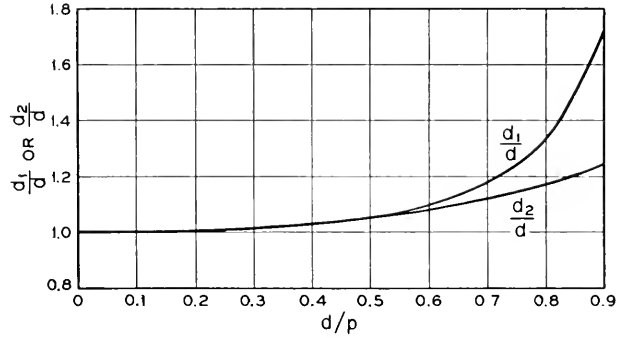

Fig. 3.16—Ratios of the wire diameters for the four turns per wavelength analysis.

The ratios R and the impedance are obtained merely by comparing the power flow for the developed sheet with a single sinusoidally distributed component with the power flow for case II for the same distant field. In a comparison with the helically conducting sheet, $n = 2$ is used in (3.50). The results are shown in Figs. 3.13, 3.14, 3.15. We see that on the basis of the largest available field, the best wire size is $d/p = .19$.

3.4 Transmission Line Equations and Helices

It is of course possible at any frequency to construct a transmission line with a distributed shunt susceptance B per unit length and a distributed shunt reactance X per unit length and, by adjusting B and X to make the phase velocity and $E^2/\beta^2 P$ the same for the artificial line as for the helix. In simulating the helix with the line, B and X must be changed as frequency is changed. Indeed, it may be necessary to change B and X somewhat in simulating a helix with a forced wave on it, as, the wave forced by an electron stream. Nevertheless, a qualitative insight into some problems can be obtained by use of this type of circuit analogue.

3.4a *Effect of Dielectric on Helix Impedance Parameter*

One possible application of the transmission line equivalent is in estimating the lowering of the helix impedance parameter $(E^2/\beta^2 P)^{1/3}$.

In the case of a transmission line of susceptance B and reactance X per unit length, we have for the phase constant β and the characteristic impedance K

$$\beta = \sqrt{BX} \tag{3.61}$$

$$K = \sqrt{X/B} \tag{3.62}$$

Now, suppose that B is increased by capacitive loading so that β has a larger value β_d. Then we see that K will have a value K_d

$$K_d = (\beta/\beta_d)K \tag{3.63}$$

Where should K be measured? It is reasonable to take the field at the surface of the helix or the helically conducting sheet as the point at which the field should be evaluated. The field at the axis will, then, be changed by a different amount, for the field at the surface of the helix is $I_0(\gamma a)$ times the field at the axis.

Suppose, then, we design a helix to have a phase constant β (a phase velocity ω/β) and, in building it, find that the dielectric supports increase the phase constant to a value β_d giving a smaller phase velocity ω/β_d. Suppose β/β_0 is large, so that γ is nearly equal to β. How will we estimate the actual axial value of $(E^2/\beta^2 P)^{1/3}$? We make the following estimate:

$$(E^2/\beta^2 P)_d^{1/3} = \left(\frac{\beta}{\beta_d}\right)^{1/3} \left(\frac{I_0(\beta a)}{I_0(\beta_d a)}\right)^{2/3} (E^2/\beta^2 P)^{1/3} \tag{3.64}$$

Here the factor $(\beta/\beta_d)^{1/3}$ is concerned with the reduction of impedance measured at the helix surface, and the other factor is concerned with the greater falling-off of the field toward the center of the helix because of the larger value of γ (taken equal to β and β_d in the two cases).

The writer does not know how good this estimate may be.

3.4b *Coupled Helices*

Another case in which the equivalent transmission line approach is particularly useful is in considering the problem of concentric helices. Such configurations have been particularly suggested for producing slow transverse fields. They can be analyzed in terms of helically conducting cylinders or in terms of developed cylinders. A certain insight can be gained very quickly, however, by the approach indicated above.

We will simulate the helices by two transmission lines of series impedances jX_1 and jX_2, of shunt admittances jB_1 and jB_2 coupled by series mutual

impedance and shunt mutual admittance jX_{12} and jB_{12}. If we consider a wave which varies as $\exp(-j\Gamma z)$ in the z direction we have

$$\Gamma I_1 - jB_1 V_1 - jB_{12} V_2 = 0 \tag{3.65}$$

$$\Gamma V_1 - jX_1 I_1 - jX_{12} I_2 = 0 \tag{3.66}$$

$$\Gamma I_2 - jB_2 V_2 - jB_{12} V_1 = 0 \tag{3.67}$$

$$\Gamma V_2 - jX_2 I_2 - jX_{12} I_1 = 0 \tag{3.68}$$

If we solve (3.65) and (3.67) for I_1 and I_2 and eliminate these, we obtain

$$\frac{V_2}{V_1} = \frac{-(\Gamma^2 + X_1 B_1 + X_{12} B_{12})}{X_1 B_{12} + B_2 X_{12}} \tag{3.69}$$

$$\frac{V_1}{V_2} = \frac{-(\Gamma^2 + X_2 B_2 + X_{12} B_{12})}{X_2 B_{12} + B_1 X_{12}} \tag{3.70}$$

Multiplying these together we obtain

$$\begin{aligned} &\Gamma^4 + (X_1 B_1 + X_2 B_2 + 2X_{12} B_{12})\Gamma^2 \\ &\quad + (X_1 X_2 - X_{12}^2)(B_1 B_2 - B_{12}^2) = 0 \end{aligned} \tag{3.71}$$

We can solve this for the two values of Γ^2

$$\begin{aligned} \Gamma^2 = &-\tfrac{1}{2}(X_1 B_1 + X_2 B_2 + 2X_{12} B_{12}) \\ &\pm \tfrac{1}{2}\left[(X_1 B_1 - X_2 B_2)^2 + 4(X_1 B_1 + X_2 B_2)(X_{12} B_{12})\right. \\ &\left. + 4(X_1 X_2 B_{12}^2 + B_1 B_2 X_{12}^2)\right]^{1/2} \end{aligned} \tag{3.72}$$

Each value of Γ^2 represents a normal mode of propagation involving both transmission lines. The two square roots of each Γ^2 of course indicate waves going in the positive and negative directions.

Suppose we substitute (3.72) into (3.69). We obtain

$$\frac{V_2}{V_1} = \frac{-(X_1 B_1 - X_2 B_2) \pm [(X_1 B_1 - X_2 B_2)^2 + 4(X_1 B_1 + X_2 B_2)(X_{12} B_{12}) + 4(X_1 X_2 B_{12}^2 + B_1 B_2 X_{12}^2)]^{1/2}}{2(X_1 B_{12} + B_2 X_{12})} \tag{3.73}$$

We will be interested in cases in which $X_1 B_1$ is very nearly equal to $X_2 B_2$. Let

$$\Delta\Gamma_0^2 = X_1 B_1 - X_2 B_2 \tag{3.74}$$

and in the parts of (3.73) where the difference of (3.74) does not occur use

$$\begin{aligned} X_1 &= X_2 = X \\ B_1 &= B_2 = B \end{aligned} \tag{3.75}$$

Then, approximately

$$\frac{V_2}{V_1} = \frac{-\Delta\Gamma_0^2 \pm [(\Delta\Gamma_0^2)^2 + 4(XB_{12} + BX_{12})^2]^{1/2}}{2(XB_{12} + BX_{12})} \tag{3.76}$$

Let us assume that $\Delta\Gamma^2$ is very small and retains terms up to the first power of $\Delta\Gamma^2$

$$\frac{V_2}{V_1} = \pm 1 + \frac{\Delta\Gamma_0^2}{2(XB_{12} + BX_{12})} \tag{3.77}$$

Let

$$\Gamma_0^2 = -XB \tag{3.78}$$

$$\frac{V_2}{V_1} = \pm 1 - \frac{\Delta\Gamma_0^2/\Gamma_0^2}{2(B_{12}/B + X_{12}/X)} \tag{3.79}$$

Let us now interpret (3.79). This says that if $\Delta\Gamma_0^2$ is zero, that is, if $X_1B_1 = X_2B_2$ exactly, there will be two modes of transmission, a *longitudinal* mode in which $V_2/V_1 = +1$ and a *transverse* mode in which $V_2/V_1 = -1$. If we excite the transverse mode it will persist. However, if $\Delta\Gamma_0^2 \neq 0$, there will be two modes, one for which $V_2 > V_1$ and the other for which $V_2 < V_1$; in other words, as $\Delta\Gamma_0^2$ is increased, we approach a condition in which one mode is nearly propagated on one helix only and the other mode nearly propagated on the other helix only. Then if we drive the pair with a transverse field we will excite both modes, and they will travel with different speeds down the system.

We see that to get a good transverse field we must make

$$\frac{\Delta\Gamma_0^2}{\Gamma_0^2} \ll 2(B_{12}/B + X_{12}/X) \tag{3.80}$$

In other words, the stronger the coupling (B_{12}, X_{12}) the more the helices can afford to differ (perhaps accidentally) in propagation constant and the pair still give a distinct transverse wave.

Thus, it seems desirable to couple the helices together as tightly as possible and especially to see that B_{12} and X_{12} have the same signs.

Let us consider two concentric helices wound in opposite directions, as in Fig. 3.17. A positive voltage V_1 will put a positive charge on helix 1 while a positive voltage V_2 will put a negative charge on helix 1. Thus, B_{12} B is negative. It is also clear that the positive current I_2 will produce flux linking helix 1 in the opposite direction from the positive current I_1, thus making X_{12}/X negative. This makes it clear that to get a good transverse field between concentric helices, the helices should be wound in opposite direc-

tions. If the helices were wound in the same direction, the "transverse" and "longitudinal" modes would cease to be clearly transverse and longitudinal should the phase velocities of the two helices by accident differ a little. Further, even if the phase velocities were the same, the transverse and longitudinal modes would have almost the same phase velocity, which in itself may be undesirable.

Field analyses of coupled helices confirm these general conclusions.

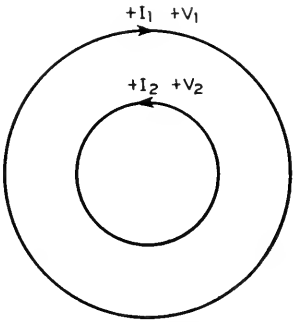

Fig. 3.17—Currents and voltages of concentric helices.

3.5 About Loss in Helices

The loss of helices is not calculated in this book. Some matters concerning deliberately added loss will be considered, however.

Loss is added to helices so that the backward loss of the tube (loss for a wave traveling from output to input) will be greater than the forward gain. If the forward gain is greater than the backward loss, the tube may oscillate if it is not terminated at each end in a good broad-band match.

In some early tubes, loss was added by making the helix out of lossy wire, such as nichrome or even iron, which is much lossier at microwave frequencies because of its ferromagnetism. Most substances are in many cases not lossy enough. Iron is very lossy, but its presence upsets magnetic focusing.

When the helix is supported by a surrounding glass tube or by parallel ceramic or glass rods, loss may be added by spraying aquadag on the inside or outside of the glass tube or on the supporting rods. This is advantageous in that the distribution of loss with distance can be controlled.

It is obvious that for lossy material a finite distance from the helix there is a resistivity which gives maximum attenuation. A perfect conductor would introduce no dissipation and neither would a perfect insulator.

If lossy material is placed a little away from the helix, loss can be made greater at lower frequencies (at which the field of the helix extends out into the lossy material) than at higher frequencies (at which the fields of

the helix are crowded near the helix and do not give rise to much current in the lossy material. This construction may be useful in preventing high-frequency tubes from oscillating at low frequencies.

Loss may be added by means of tubes or collars of lossy ceramic which fit around the helix.

APPENDIX I

MISCELLANEOUS INFORMATION

This appendix presents an assortment of material which may be useful to the reader.

Constants

Electronic charge-to-mass ratio:

$\eta = e/m = 1.759 \times 10^{11}$ Coulomb/kilogram

Electronic charge: $e = 1.602 \times 10^{-19}$ Coulomb

Dielectric constant of vacuum: $\epsilon = 8.854 \times 10^{-12}$ Coulomb/meter

Permitivity of vacuum: $\mu = 1.257 \times 10^{-6}$ Henry/meter

Boltzman's constant: $k = 1.380 \times 10^{-23}$ Joule/degree

Cross Products

$$(A' \times A'')_x = A_y' A_z'' - A_z' A_y''$$
$$(A' \times A'')_y = A_z' A_x'' - A_x' A_z''$$
$$(A' \times A'')_z = A_x' A_y'' - A_y' A_x''$$

Maxwell's Equations: Rectangular Coordinates

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = -j\omega\mu H_x \qquad \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = j\omega\epsilon E_x + J_x$$

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = -j\omega\mu H_y \qquad \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = j\omega\epsilon E_y + J_y$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -j\omega\mu H_z \qquad \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = j\omega\epsilon E_z + J_z$$

Maxwell's Equations: Axially Symmetrical

$$\frac{\partial E_\varphi}{\partial z} = -j\omega\mu H_\rho \qquad \frac{\partial H_\varphi}{\partial z} = -(j\omega\epsilon E_\rho + J_\rho)$$

$$\frac{\partial E_\rho}{\partial z} - \frac{\partial E_z}{\partial \rho} = -j\omega\mu H_\varphi \qquad \frac{\partial H_\rho}{\partial z} - \frac{\partial H_z}{\partial \rho} = j\omega\epsilon E_\varphi + J_\varphi$$

$$\frac{\partial}{\partial \rho}(\rho E_\varphi) = -j\omega\mu\rho H_z \qquad \frac{\partial}{\partial \rho}(\rho H_\varphi) = \rho(j\omega\epsilon E_z + J_z)$$

MISCELLANEOUS FORMULAE INVOLVING $I_n(x)$ AND $K_n(x)$

1. $I_{\nu-1}(Z) - I_{\nu+1}(Z) = \frac{2\nu}{Z} I_\nu(Z), \qquad K_{\nu-1}(Z) - K_{\nu+1}(Z) = -\frac{2\nu}{Z} K_\nu(Z)$

2. $I_{\nu-1}(Z) + I_{\nu+1}(Z) = 2I'_\nu(Z), \qquad K_{\nu-1}(Z) + K_{\nu+1}(Z) = -2K'_\nu(Z)$

3. $ZI'_\nu(Z) + \nu I_\nu(Z) = ZI_{\nu-1}(Z), \qquad ZK'_\nu(Z) + \nu K_\nu(Z) = -ZK_{\nu-1}(Z)$

4. $ZI'_\nu(Z) - \nu I_\nu(Z) = ZI_{\nu+1}(Z), \qquad ZK'_\nu(Z) - \nu K_\nu(Z) = -ZK_{\nu+1}(Z)$

5. $\left(\frac{d}{ZDZ}\right)^m \{Z^\nu I_\nu(Z)\} = Z^{\nu-m} I_{\nu-m}(Z), \left(\frac{d}{ZdZ}\right)^m \{Z^\nu K_\nu(Z)\}$

$$= (-)^m Z^{\nu-m} K_{\nu-m}(Z)$$

6. $\left(\frac{d}{ZdZ}\right)^m \left\{\frac{I_\nu(Z)}{Z^\nu}\right\} = \frac{I_{\nu+m}(Z)}{Z^{\nu+m}}, \qquad \left(\frac{d}{ZdZ}\right)^m \left\{\frac{K_\nu(Z)}{Z^\nu}\right\} = (-)^m \frac{K_{\nu+m}(Z)}{Z^{\nu+m}}$

7. $I'_0(Z) = I_1(Z), \qquad K'_0(Z) = -K_1(Z)$

8. $I_{-\nu}(Z) = I_\nu(Z), \qquad K_{-\nu}(Z) = K_\nu(Z)$

9. $K_{1/2}(Z) = \left(\frac{\pi}{2Z}\right)^{1/2} e^{-z}$

10. $I_\nu(Ze^{m\pi i}) = e^{m\nu\pi i} I_\nu(Z)$

11. $K_\nu(Ze^{m\pi i}) = e^{-m\nu\pi i} K_\nu(Z) - i \frac{\sin m\nu\pi}{\sin \nu\pi} I_\nu(Z)$

12. $I_\nu(Z) K_{\nu+1}(Z) + I_{\nu+1}(Z) K_\nu(Z) = 1/Z$

For small values of X:

13. $I_0(X) = 1 + .25 X^2 + .015625 X^4 + \cdots$

14. $I_1(X) = .5X + .0625 X^3 + .002604 X^5 + \cdots$

15. $K_0(X) = -\left\{\gamma + \ln\left(\frac{X}{2}\right)\right\} I_0(X) + \frac{1}{4} X^2 + \frac{3}{128} X^4 + \cdots\Big\}$

16. $K_1(X) = \left\{\gamma + \ln\left(\frac{X}{2}\right)\right\} I_1(X) + \frac{1}{X} - \frac{1}{4} X - \frac{5}{64} X^3 + \cdots\Big\}$

$$\gamma = .5772 \ldots \text{ (Euler's constant)}$$

For large values of X:

17. $I_0(X) \sim \frac{e^X}{(2\pi X)^{1/2}} \left\{1 + \frac{.125}{X} + \frac{.0703125}{X^2} + \frac{.073242}{X^3} + \cdots\right\}$

$$18.\ I_1(X) \sim \frac{e^X}{(2\pi X)^{1/2}}\left\{1 - \frac{.375}{X} - \frac{.1171875}{X^2} - \frac{.102539}{X^3} - \cdots\right\}$$

$$19.\ K_0(X) \sim \left(\frac{\pi}{2X}\right)^{1/2} e^{-X}\left\{1 - \frac{.125}{X} + \frac{.0703125}{X^2} - \frac{.073242}{X^3} + \cdots\right\}$$

$$20.\ K_1(X) \sim \left(\frac{\pi}{2X}\right)^{1/2} e^{-X}\left\{1 + \frac{.375}{X} - \frac{.1171875}{X^2} + \frac{.102539}{X^3} - \cdots\right\}.$$

Fig. A1.1 shows $I_0(X)$ (solid line) and the first two terms of 13 and the first term of 17 (dashed lines).

Fig. A1.2 shows $I_1(X)$ (solid line) and the first term of 14 and the first term of 18 (dashed lines).

Fig. A1.3 shows $K_0(X)$ (solid line) and $-\left\{\gamma + \ln\left(\frac{X}{2}\right)\right\} I_0(X)$ and the first term of 19 (dashed lines).

Fig. A1.4 shows $K_1(X)$ (solid line) and $\left\{\gamma + \ln\left(\frac{X}{2}\right)\right\} I_1(X) + 1/X$ and the first term of 20 (dashed lines).

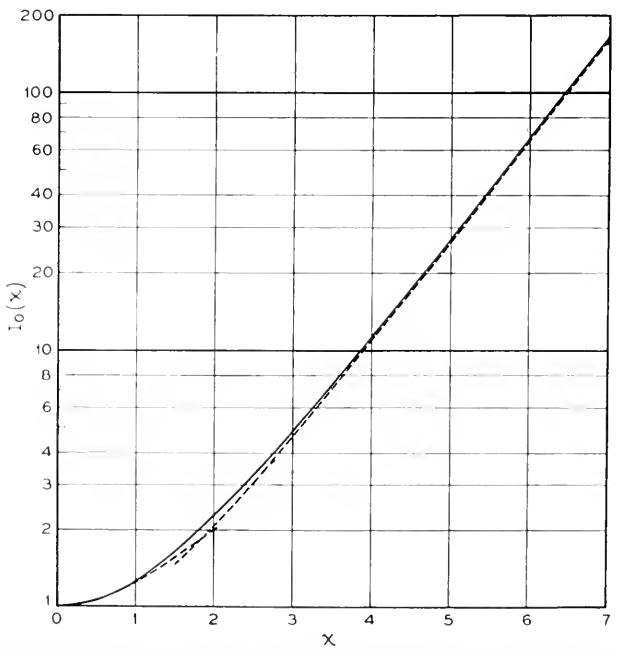

Fig. A1.1 — The correct value of $I_0(X)$ (solid line), the first two terms of the series expansion 13 (dashed line from origin), and the first term of the asymptotic series 17 (dashed line to right)

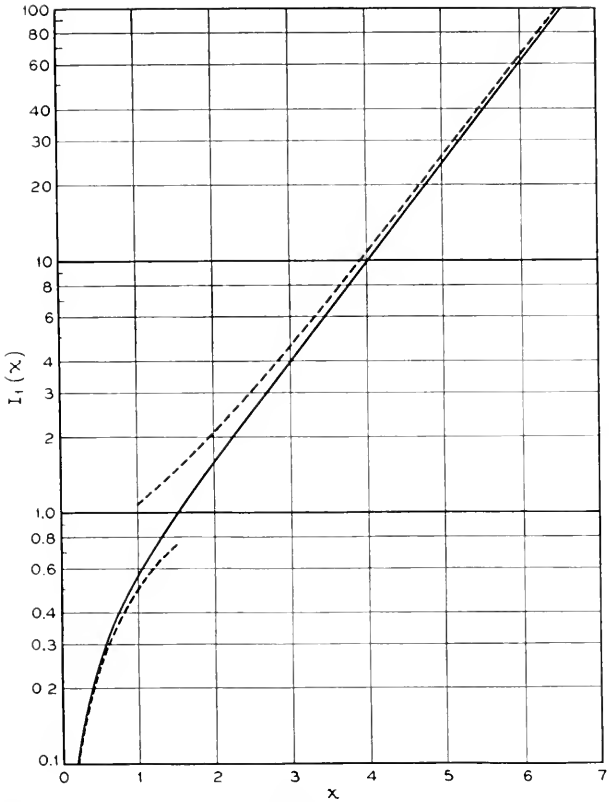

Fig. A1.2—The correct value of $I_1(X)$ (solid line), the first term of the series expansion 14 (lower dashed line), and the first term of the asymptotic series 18 (upper dashed line).

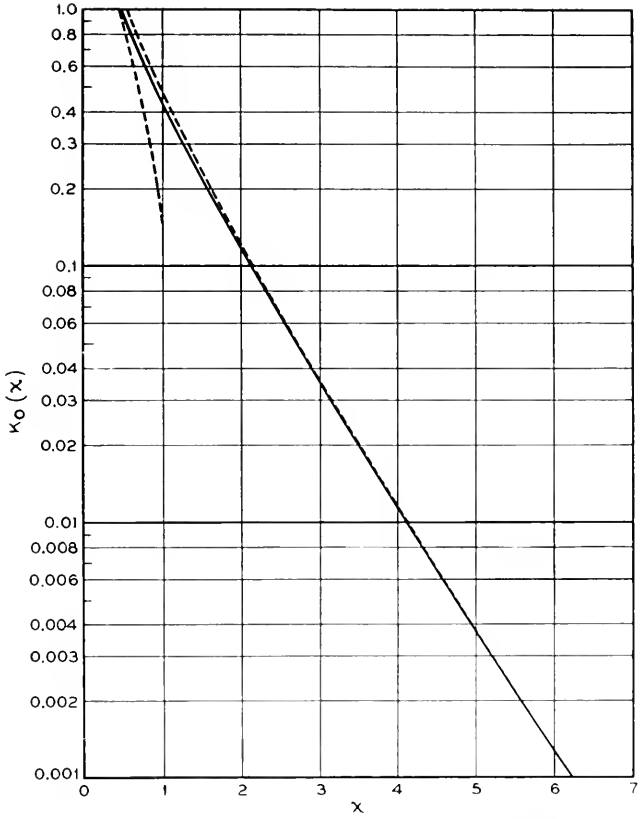

Fig. A1.3—The correct value of $K_0(X)$ (solid line), $-\left\{\gamma + \ln\left(\frac{X}{2}\right)\right\} I_0(X)$ from the series expansion 15 (left dashed line), and the first term of the asymptotic series 19 (right dashed line).

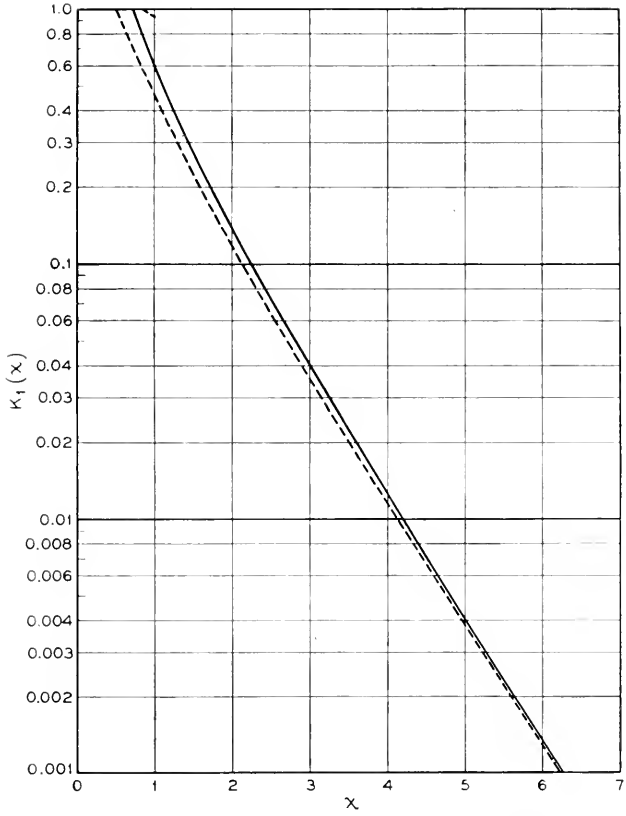

Fig. A1.4—The correct value of $K_1(X)$ (solid line), $\left\{\gamma + \ln\left(\frac{X}{2}\right)\right\} I_1(X)$ from the series expansion 16 (upper dashed line), and the first term of the asymptotic series 20 (lower dashed line).

APPENDIX II

PROPAGATION ON A HELICALLY CONDUCTING CYLINDER

The circuit parameter important in the operation of traveling-wave tubes is:

$$(E_z^2/\beta^2 P)^{1/3} \tag{1}$$

$$\beta = \omega/v. \tag{2}$$

Here E_z is the peak electric field in the direction of propagation, P is the power flow along the helix, and v is the phase velocity of the wave. The quantity $E_z^2/\beta^2 P$ has the dimensions of impedance.

While the problem of propagation along a helix has not been solved, what appears to be a very good approximation has been obtained by replacing the helix with a cylinder of the same mean radius a which is conducting only in a helical direction making an angle Ψ with the circumference, and nonconducting in the helical direction normal to this.

An appropriate solution of the wave equation in cylindrical co-ordinates for a plane wave having circular symmetry and propagating in the z direction with velocity

$$v = \frac{\omega}{\beta}, \tag{3}$$

less than the speed of light c, is

$$E_z = [AI_0(\gamma r) + BK_0(\gamma r)]e^{j(\omega t - \beta_z)} \tag{4}$$

where I_0 and K_0 are the modified Bessel functions, and

$$\gamma^2 = \beta^2 - \left(\frac{\omega}{c}\right)^2 = \beta^2 - \beta_0^2. \tag{5}$$

The form of the z (longitudinal) components of an electromagnetic field varying as $e^{j(\omega t - \beta z)}$ and remaining everywhere finite might therefore be

$$H_{z1} = B_1 I_0(\gamma r)e^{j(\omega t - \beta z)} \tag{6}$$

$$E_{z3} = B_3 I_0(\gamma r)e^{j(\omega t - \beta z)} \tag{7}$$

inside radius a, and

$$H_{z2} = B_2 K_0(\gamma r)e^{j(\omega t - \beta z)} \tag{8}$$

$$E_{z4} = B_4 K_0(\gamma r) e^{j(\omega t - \beta z)} \tag{9}$$

outside radius α. Omitting the factor $e^{j(\omega t-\beta z)}$ the radial and circumferential components associated with these, obtained by applying the curl equation, are, inside radius α,

$$H_{\phi 3} = B_3 \frac{j\omega\epsilon}{\gamma} I_1(\gamma r) \tag{10}$$

$$H_{r1} = B_1 \frac{j\beta}{\gamma} I_1(\gamma r) \tag{11}$$

$$E_{\phi 1} = -B_1 \frac{j\omega\mu}{\gamma} I_1(\gamma r) \tag{12}$$

$$E_{r3} = B_3 \frac{j\beta}{\gamma} I_1(\gamma r) \tag{13}$$

and outside radius α

$$H_{\phi 4} = -B_4 \frac{j\omega\epsilon}{\gamma} K_1(\gamma r) \tag{14}$$

$$H_{r2} = -B_2 \frac{j\beta}{\gamma} K_1(\gamma r) \tag{15}$$

$$E_{\phi 3} = B_2 \frac{j\omega\mu}{\gamma} K_1(\gamma r) \tag{16}$$

$$E_{r4} = -B_4 \frac{j\beta}{\gamma} K_1(\gamma r). \tag{17}$$

The boundary conditions which must be satisfied at the cylinder of radius α are that the tangential electric field must be perpendicular to the helix direction

$$E_{z3} \sin \Psi + E_{\phi 1} \cos \Psi = 0 \tag{18}$$

$$E_{z4} \sin \Psi + E_{\phi 2} \cos \Psi = 0, \tag{19}$$

the tangential electric field must be continuous across the cylinder

$$E_{z3} = E_{z4} \text{ (and } E_{\phi 1} = E_{\phi 2}), \tag{20}$$

and the tangential component of magnetic field parallel to the helix direction must be continuous across the cylinder, since there can be no current in the surface perpendicular to this direction.

$$H_{z1} \sin \Psi + H_{\phi 3} \cos \Psi = H_{z2} \sin \Psi + H_{\phi 4} \cos \Psi. \tag{21}$$

These equations serve to determine the ratios of the B's and to determine γ through

$$(\gamma\alpha)^2 \frac{I_0(\gamma\alpha)K_0(\gamma\alpha)}{I_1(\gamma\alpha)K_1(\gamma\alpha)} = (\beta_0 \alpha \cot \Psi)^2. \tag{22}$$

We can easily express the various field components listed in (6) through (17) in terms of a common amplitude factor. As such expressions are useful in understanding the nature of the field, it seems desirable to list them in an orderly fashion.

Inside the Helix:

$$E_z = BI_0(\gamma r)e^{j(\omega t-\beta z)} \tag{23}$$

$$E_r = j\beta \frac{\beta}{\gamma} I_1(\gamma r)e^{j(\omega t-\beta z)} \tag{24}$$

$$E_\Phi = -B \frac{I_0(\gamma a)}{I_1(\gamma a)} \frac{1}{\cot \psi} I_1(\gamma r)e^{j(\omega t-\beta z)} \tag{25}$$

$$H_z = -j \frac{B}{k} \frac{\gamma}{\beta_0} \frac{I(\gamma a)}{I_1(\gamma a)} \frac{1}{\cot \psi} I_0(\gamma r)e^{j(\omega t-\beta z)} \tag{26}$$

$$H_r = \frac{B}{k} \frac{\beta}{\beta_0} \frac{I_0(\gamma a)}{I_1(\gamma a)} \frac{1}{\cot \psi} I_1(\gamma r)e^{j(\omega t-\beta z)} \tag{27}$$

$$H_\Phi = j \frac{B}{k} \frac{\beta_0}{\gamma} I(\gamma r)e^{j(\omega t-\beta z)}. \tag{28}$$

Outside the Helix:

$$E_z = B \frac{I_0(\gamma a)}{K_0(\gamma a)} K_0(\gamma r)e^{j(\omega t-\beta z)} \tag{29}$$

$$E_r = -jB \frac{\beta}{\gamma} \frac{I_0(\gamma a)}{K_0(\gamma a)} K_1(\gamma r)e^{j(\omega t-\beta z)} \tag{30}$$

$$E_\Phi = -B \frac{I_0(\gamma a)}{K_1(\gamma a)} \frac{1}{\cot \psi} K_1(\gamma r)e^{j(\omega t-\beta z)} \tag{31}$$

$$H_z = j \frac{B}{k} \frac{\gamma}{\beta_0} \frac{I_0(\gamma a)}{K_1(\gamma a)} \frac{1}{\cot \psi} K_0(\gamma r)e^{j(\omega t-\beta z)} \tag{32}$$

$$H_r = \frac{B}{k} \frac{I_0(\gamma a)}{K_1(\gamma a)} \frac{1}{\cot \psi} K_1(\gamma r)e^{j(\omega t-\beta z)} \tag{33}$$

$$H_\Phi = -j \frac{B}{k} \frac{\beta_0}{\gamma} \frac{I_0(\gamma a)}{K_0(\gamma a)} K_1(\gamma r)e^{j(\omega t-\beta z)} \tag{34}$$

Here

$$k = \sqrt{\mu/\epsilon} = 120\,\pi \text{ ohms} \tag{35}$$

The power associated with the propagation is given by

$$P = \tfrac{1}{2} \operatorname{Re} \int E \times H^* \, d\tau \tag{36}$$

taken over a plane normal to the axis of propagation. This is

$$P = \pi \operatorname{Re} \left[\int_0^a (E_r H_\Phi^* - E_\Phi H_r^*) r \, dr + \int_a^\infty (E_r H_\Phi^* - E_\Phi H_r^*) r \, dr \right] \tag{37}$$

or

$$\begin{aligned} P &= \pi E_z^2(0) \frac{\beta \beta_0^2}{\gamma^2 \omega \mu} \left[\left(1 + c \frac{I_0 K_1}{I_1 K_0} \right) \int_0^a I_1^2(\gamma r) r \, dr \right. \\ &\qquad \left. + \left(\frac{I_0}{K_0} \right)^2 \left(1 + \frac{I_1 K_0}{I_0 K_1} \right) \int_a^\infty K_1^2(\gamma r) r \, dr \right] \\ &= E_z^2(0) \frac{\pi}{2k} \frac{\beta \beta_0 a^2}{\gamma^2} \left[\left(1 + \frac{I_0 K_1}{I_1 K_0} \right) (I_1^2 - I_0 I_2) \right. \\ &\qquad \left. + \left(\frac{I_0}{K_0} \right)^2 \left(1 + \frac{I_1 K_0}{I_0 K_1} \right) (K_0 K_2 - K_1^2) \right]. \end{aligned} \tag{38}$$

where $k = 120 \pi$ ohms.
Let us now write

$$(E_z^2/\beta^2 P)^{1/3} = (\beta/\beta_0)^{1/3} (\gamma/\beta)^{4/3} F(\gamma a) \tag{39}$$

where

$$\begin{aligned} F(\gamma\alpha) = \left\{ \left(\frac{(\gamma\alpha)^2}{240} \right) \left[(I_1^2 - I_0 I_2) \left(1 + \frac{I_0 K_1}{I_1 K_0} \right) \right. \right. \\ \left. \left. + \left(\frac{I_0}{K_0} \right)^2 (K_0 K_2 - K_1^2) \left(1 + \frac{I_1 K_0}{K_1 I_0} \right) \right] \right\}^{-1/3}. \end{aligned} \tag{40}$$

We can rewrite the expression for $F(\gamma\alpha)$ by using relations, Appendix I:

$$F(\gamma\alpha) = \left(\frac{\gamma\alpha}{240} \frac{I_0}{K_0} \left[\left(\frac{I_1}{I_0} - \frac{I_0}{I_1} \right) + \left(\frac{K_0}{K_1} - \frac{K_1}{K_0} \right) + \frac{4}{\gamma\alpha} \right] \right)^{-1/3}. \tag{41}$$

Communication in the Presence of Noise—Probability of Error for Two Encoding Schemes

By S. O. RICE

Recent work by C. E. Shannon and others has led to an expression for the maximum rate at which information can be transmitted in the presence of random noise. Here two encoding schemes are described in which the ideal rate is approached when the signal length is increased. Both schemes are based upon drawing random numbers from a normal universe, an idea suggested by Shannon's observation that in an efficient encoding system the typical signal will resemble random noise. In choosing these schemes two requirements were kept in mind: (1) the ideal rate must be approached, and (2) the problem of computing the probability of error must be tractable. Although both schemes meet both requirements, considerable work has been required to put the expression for the probability of error into manageable form.

1. Introduction

In recent work concerning the theory of communication it has been shown that the maximum or ideal rate of signaling which may be achieved in the presence of noise is (1, 2, 3, 4, 5)

$$R_I = F \log_2 (1 + W_S/W_N) \text{ bits/sec.} \tag{1-1}$$

In this expression F is the width of the frequency band used for signaling (which we suppose to extend from 0 to F cps), W_S is the average signaling power and W_N the average power of the noise. The noise is assumed to be random and to have a constant power spectrum of W_N/F watts per cps over the frequency band $(0, F)$.

This ideal rate is achieved only by the most efficient encoding schemes in which, as Shannon (1, 2) states, the typical signal has many of the properties of random noise. Here we shall study two different encoding schemes, both of them referring to a bandwidth F and a time interval T. By making the product FT large enough the ideal rate of signaling may be approached in either case* and we are interested in the probability of error for rates of signaling a little below the rate (1-1). The work given here is closely associated with Section 7 of Shannon's second paper (2).

In the first encoding scheme the signal corresponding to a given message lasts exactly T seconds, but (because the signal is zero outside this assigned interval of duration) the power spectrum of the signal is not exactly zero for frequencies exceeding F. In the second encoding scheme, the signal

* A recent analysis by M. J. E. Golay (*Proc. I. R. E.*, Sept. 1949, p. 1031) indicates that the ideal rate of signaling may also be approached by quantized PPM under suitable conditions.

power spectrum is limited to the band $(0, F)$ but the signal, regarded as a function of time, is not exactly zero outside its allotted interval of length T.

It turns out that both schemes lead to the same mathematical problem which may be stated as follows: Given two universes of random numbers both distributed normally about zero with standard deviations σ and ν, respectively. Let the first universe be called the σ (signal) universe and the second the ν (noise) universe. Draw $2N + 1$ numbers $A_{-N}^{(0)}, A_{-N+1}^{(0)}, \cdots, A_0^{(0)}, \cdots, A_N^{(0)}$ at random from the σ universe. These $2N + 1$ numbers may be regarded as the rectangular coordinates of a point P_0 in $2N + 1$-dimensional space. Draw $2N + 1$ numbers $B_{-N}, \cdots, B_0, \cdots, B_N$ at random from the ν universe and imagine a (hyper-) sphere S of radius $x_0^{1/2} = P_0Q$, where

$$x_0 = \sum_{n=-N}^{N} B_n^2 = \overline{P_0Q}^2, \tag{1-2}$$

centered on the point Q whose coordinates are $A_n^{(0)} + B_n$, $n = -N, \cdots, 0, \cdots, N$. Return to the σ universe, draw out K sets of $2N + 1$ numbers each, denote the kth set by $A_{-N}^{(k)}, \cdots, A_0^{(k)}, \cdots, A_N^{(k)}$ and the associated point by P_k.

What is the probability that none of the K points $P_1, \cdots, P_K$ lie within the sphere S? In other words what is the probability, which will be denoted by "Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$," that the K distances $P_1Q, \cdots, P_KQ$ will all exceed the radius P_0Q? In terms of the A_n's and B_n's we ask for the probability that all K of the numbers $x_1, x_2, \cdots, x_K$ exceed x_0 where

$$x_k = \sum_{n=-N}^{N} (A_n^{(k)} - A_n^{(0)} - B_n)^2 = \overline{P_kQ}^2 \tag{1-3}$$

Expression (1-2) for x_0 is seen to be a special case of (1-3). The relationship between the points P_0, Q, P_1, $P_2, \cdots, P_k, \cdots, P_K$ is indicated in Fig. 1.

The answer to this problem is given by the rather complicated expression (4-12) which, when written out, involves Bessel functions of imaginary argument and of order $N - 1/2$. When N and K become very large the work of Section 5 shows that the probability in question is given by

$$\text{Prob. } (P_1Q, \cdots, P_KQ > P_0Q) = (1 + \operatorname{erf} H)/2 + 0(1/K) + 0(N^{-1/2} \log^{3/2} N) \tag{1-4}$$

where, with $r = \nu^2/\sigma^2$,

$$H = \left(\frac{1 + r}{4N}\right)^{1/2} \left[(N + 1/2) \log_e (1 + 1/r) - \log_e (K + 1) + \frac{1}{2} \log_e \frac{2\pi N(1 + 2r)}{(1 + r)^2} \right] \tag{1-5}$$

The symbol $0(N^{-1/2} \log^{3/2} N)$ stands for a term of order $N^{-1/2} \log^{3/2} N$, i.e., a positive constant C and a value N_0 can be found such that the absolute value of the term in question is less than $CN^{-1/2} \log^{3/2} N$ when $N > N_0$. In order to obtain actual numerical values for C and N_0, considerably more work than is given here would be required. The term $0(1/K)$ is of the same nature. The "order of" terms have been carried along in the work of Section 5 in order to guard against error in the many approximations which are made in the derivation of (1-4).

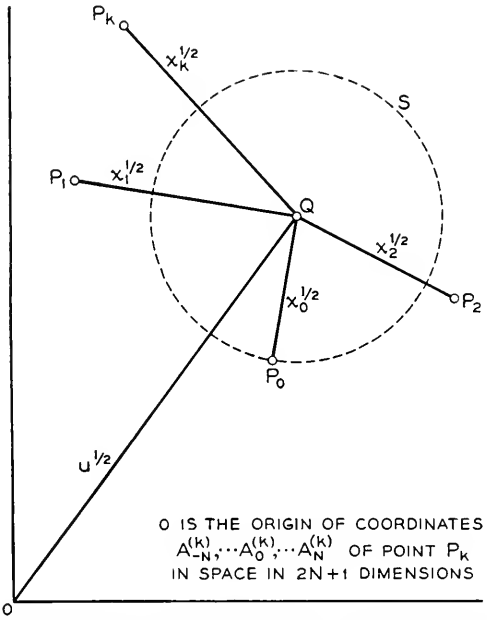

Fig. 1—Diagram indicating relationship between points P_0, Q, and P_k corresponding to signal, signal plus noise, and k^{th} signal not sent ($k > 0$), respectively.

The last term within the bracket in (1-5) has been retained even though it gives terms of order $N^{-1/2} \log N$ when (1-5) is put in (1-4) and could thus be included in $0(N^{-1/2} \log^{3/2} N)$. As shown by the table in the next paragraph, inclusion of this term considerably improves the agreement between (1-4) and values of Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$ obtained by integrating the exact expression (4-12) numerically. This suggests that the term $0(N^{-1/2} \log^{3/2} N)$ in (1-4) is unnecessarily large.

Although the "order of" terms in (1-4) give us some idea of the accuracy of the approximation expressed by (1-4) and (1-5), a better one is desirable. With this in mind the lengthy task of computing the exact expression (4-12) for Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$ by numerical integration was undertaken.

The values obtained in this way are listed in the second column of the following table. The values of Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$ obtained from (1-4) (in which the "order of" terms are ignored) and (1-5) are given in the third column. Column IV lists values obtained from (1-4) and a simplified form of (1-5) obtained by omitting the last term in (1-5). These values are less accurate than those in the third column. The values in Column V are computed from (1-5) and a modified form of (1-4) obtained by adding the correction term shown in equation (5-53) (with $B = H$). The values in Column V are presumably the best that can be done with the approximations made in Section V of this paper, although the first entry renders this a little doubtful.

Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$ for $N = 99.5$ & $r = 1$

$K + 1$	*Numerical Integration*	(1-4) & (1-5)	*Col. IV*	*Col. V*
$2^{100}e^{-30}$	.994	.9995	.9987	1.0001
$2^{100}e^{-15}$	.962	.9650	.9337	.9710
2^{100}	.603	.621	.5000	.605
$2^{100}e^{15}$	.1196	.1159	.0663	.1176
$2^{100}e^{30}$	.0065	.00347	.0013	.00586

It will become apparent later that the value $K + 1 = 2^{100}$ corresponds to the ideal rate of signaling. The non-integer value of 99.5 for N is explained by the fact that the calculations were started before the present version of the theory was worked out. It will be noticed that for $K + 1 = 2^{100}e^{-30}$ all of the approximate values exceed the .994 obtained by numerical integration. I am in doubt as to whether the major part of the discrepancy is due to errors in numerical integration (due to the considerable difficulty encountered) or to errors in the approximations.

In both encoding schemes, the point P_0 corresponds to the transmitted signal, Q to the transmitted signal plus noise, and P_1, P_2, $\cdots$ P_K to K other possible signals. The average signal power turns out to be $(N + 1/2)\sigma^2$ and the average noise power to be $(N + 1/2)\nu^2$. Furthermore,

x_0 = twice the average power in the noise.
x_k = " " " " " " " plus the kth signal.

Prob. $(P_1Q, \cdots P_KQ > P_0Q)$ = Probability that none of the K other signals will be mistaken for the signal sent, i.e., the probability of no error.

The random numbers $A_n^{(k)}$ are taken to be distributed normally instead of some other way because this choice makes the encoding signals (in our two schemes) resemble random noise, a condition which seems to be necessary for efficient encoding (1, 2).

Both of the encoding schemes are concerned with sending, in an interval of duration T, one of $K + 1$ different messages. According to communication theory (1, 2, 3) this corresponds to sending at the rate of $T^{-1} \log_2 (K + 1)$ bits per second. However, instead of discussing the rate of transmission, it is more convenient, from the standpoint of (1-4), to deal with the total number of bits of information sent in time T. Thus, selecting and sending one of the $K + 1$ possible messages is equivalent to sending

$$M = \log_2(K + 1) \tag{1-6}$$

bits of information. M, or one of the adjacent integers if M is not an integer, is the number of "yes or no" questions required to select the sent message from the $K + 1$ possible messages (divide the $K + 1$ messages into two equal, or nearly equal, groups; select the group containing the sent message by asking the person who knows, "Is the sent message in the first group?"; proceed in this way until the last subgroup consists of only the sent message). The amount of information which would be sent in time T at the ideal rate R_I defined by (1-1) is

$$M_I = TR_I = FT \log_2 (1 + 1/r) = (N + 1/2) \log_2 (1 + 1/r) \tag{1-7}$$

where use has been made of $W_N/W_S = \nu^2/\sigma^2 = r$, and the relation $N < FT < N + 1$ (which turns out to be common to both encoding schemes) has been approximated by $N + 1/2 = FT$.

When (1-6) and (1-7) are used to eliminate N and K from (1-5) the result is an expression for the actual amount M of information sent (in time T) in terms of (1) the amount M_I which is sent by transmitting at the ideal rate (1-1) for a time T, (2) the ratio r of the noise power to the signal power, and (3) the probability of no error in sending M bits of information in time T, this probability being given as $(1 + \text{erf } H)/2$:

$$M = M_I - aM_I^{1/2}H + b \tag{1-8}$$

where

$$a = 2\left[\frac{\log_2 e}{(1 + r) \log_e (1 + 1/r)}\right]^{1/2},$$
$$b = \frac{1}{2} \log_2 \left[\frac{2\pi(1 + 2r)M_I}{(1 + r)^2 \log_2 (1 + 1/r)}\right] \tag{1-9}$$

Here the "order of" terms in (1-4) have been neglected together with similar terms which arise when $N + 1/2$ is used for N in computing a and b. The term b is usually small compared to $aM_I^{1/2}H$.

The more slowly we send, the less chance there is of error. The relationship between M, M_I and the probability of no error, as computed from

(1–8), is shown in the following table. The probability of no error is denoted by p and the terms are given in the same order as on the right of (1–8) in order to show their relative importance. The ratio $M/M_1(=R/R_1)$ for $r = 0.1$ is shown as a function of M in Fig. 2.

For $r = W_N/W_S = 0.1$

M_1 bits	M for $p = .5$	M for $p = .99$	M for $p = .99999$
10^2	$M_1 - 0 + 3.75$	$M_1 - 24.3 + 3.75$	$M_1 - 44.6 + 3.75$
10^4	" " $+ 7.07$	" $- 243 + 7.07$	" $- 446 + 7.07$
10^6	" " $+ 10.38$	" $- 2430 + 10.38$	" $- 4460 + 10.38$

For $r = W_N/W_S = 1$

10^2	$M_1 - 0 + 4.44$	$M_1 - 33.4 + 4.44$	$M_1 - 61.2 + 4.44$
10^4	" " $+ 7.76$	" $- 334 + 7.76$	" $- 612 + 7.76$
10^6	" " $+ 11.08$	" $- 3340 + 11.08$	" $- 6120 + 11.08$

There may be some question as to the accuracy of the values for $p = .99999$, especially for $M_1 = 100$, since this corresponds to points on the tail of the probability distribution where the "order of" terms in (1-4) become relatively important.

Of course, for a given bandwidth, the ideal rate of signaling R_1 (given by (1-1)) for $r = .1$ exceeds that for $r = 1$ in the ratio $(\log_2 11)/(\log_2 2) = 3.46$.

The above results agree with the statement that, by efficient encoding, the rate of signaling R can be made to approach the ideal rate $R_1 = M_1/T$ given by (1-1). As applied to our two schemes, the term "efficient encoding" means using a very large value of FT or N. To see this, divide both sides of (1-8) by M_1 and rearrange the terms:

$$1 - M/M_1 = aH\, M_1^{-1/2} + 0(M_1^{-1} \log M_1) \qquad (1\text{-}10)$$

When M_1 is replaced by R_1T in M/M_1, the fraction M/T occurs. We shall set $R = M/T$ and call R the rate of signaling corresponding to some fixed probability of error (which determines H). Thus, when (1-7) and the definition (1-9) for a are used, (1-10) goes into

$$\frac{(R_1 - R)}{R_1} = \frac{2H}{[(1 + r)FT]^{1/2} \log_e (1 + 1/r)} + 0((\log FT)/FT) \qquad (1\text{–}11)$$

Equation (1-11) shows that when r and H are fixed (i.e. when the noise power/signal power and the probability of error are fixed) R/R_1 approaches unity as $FT \to \infty$. This is shown in Fig. 2 for the case $r = 0.1$. Since $R/R_1 = M/M_1$, M/M_1 must approach unity and consequently M as well as M_1 in-

creases linearly with FT. Thus, for efficient encoding M is large and, from (1-6), so is K.

It should be remembered that equation (1-8) has been established only for the two encoding schemes of this article. The question of how much faster M/T approaches R_I for the more efficient encoding schemes mentioned at the end of Section 2 still remains unanswered.

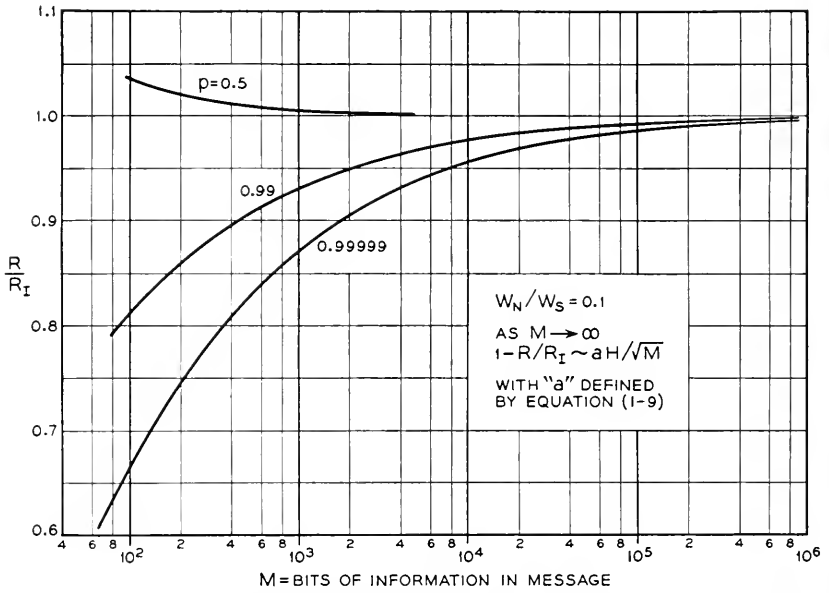

Fig. 2—Curves showing the approach of R/R_I ($= M/M_I$) to unity as the message length increases and the probability of no error remains fixed. R is the rate of signaling at which the probability of no error is p and R_I is the ideal rate.

It gives me pleasure to acknowledge the help I have received in the preparation of this memorandum from conversations with Messrs. H. Nyquist, John Riordan, C. E. Shannon, and M. K. Zinn. I am also indebted to Miss M. Darville for computing the tables shown above and for checking a number of the equations numerically.

2. The First Encoding Scheme

Suppose that we have $K + 1$ different messages any one of which is to be transmitted over a uniform frequency band extending from zero to the nominal cut-off frequency F in a time interval of length T. The adjective "nominal" is used because the sudden starting and stopping of the signals given by the first encoding scheme produces frequency components higher

than F. A shortcoming of this nature must be accepted since it is impossible to have a signal possessing both finite duration and finite bandwidth.

The first step of the encoding process is to compute the integer N given by

$$N < FT < N + 1 \tag{2-1}$$

We assume that FT is not an integer in order to avoid borderline cases. Let W_S be the average signal power available for transmission and define the standard deviation σ of the σ universe introduced in Section 1 by $(N + 1/2)\sigma^2 = W_S$. To encode the first message, draw $2N + 1$ numbers $A_{-N}^{(0)}, \cdots, A_0^{(0)}, \cdots A_N^{(0)}$ at random from the σ universe. The signal corresponding to the first message is then taken to be

$$I_0(t) = 2^{-1/2}A_0^{(0)} + \sum_{n=1}^{N} (A_n^{(0)} \cos 2\pi nt/T + A_{-n}^{(0)} \sin 2\pi nt/T) \tag{2-2}$$

The remaining K messages are encoded in the same way, the signal representing the kth message being

$$I_k(t) = 2^{-1/2}A_0^{(k)} + \sum_{n=1}^{N} (A_n^{(k)} \cos 2\pi nt/T + A_{-n}^{(k)} \sin 2\pi nt/T). \tag{2-3}$$

It is apparent that each signal consists of a d-c term plus terms corresponding to N discrete frequencies, the highest being $N/T < F$, and that the average power (assuming $I_k(t)$ to flow through a unit resistance) in the kth signal is

$$T^{-1} \int_{-T/2}^{T/2} I_k^2(t)\, dt = 2^{-1}(A_0^{(k)})^2 + \sum_{n=1}^{N} 2^{-1}[(A_n^{(k)})^2 + (A_{-n}^{(k)})^2] \tag{2-4}$$

Since the A's were drawn from a universe of standard deviation σ, the expected value of the right hand side is $(2N + 1)\sigma^2/2$ which is equal to the average signal power W_S, as required.

We pick one of the $K + 1$ messages at random and send the corresponding signal over a transmission system subject to noise. We choose our notation so that the sent signal is represented by $I_0(t)$ as given by (2-2). Let the noise be given by

$$J(t) = 2^{-1/2}B_0 + \sum_{n=1}^{N} (B_n \cos 2\pi nt/T + B_{-n} \sin 2\pi nt/T) \tag{2-5}$$

where $B_{-N}, \cdots, B_0, \cdots, B_N$ are $(2N + 1)$ numbers drawn at random from the normally distributed ν universe mentioned in the introduction. The standard deviation ν of the universe is given by $(N + 1/2)\nu^2 = W_N$, W_N being the average noise power. We call $J(t)$ simply "noise" rather than

"random noise" to emphasize that (2-5) does not represent a random noise current unless N and T approach infinity.

The input to the receiver is $I_0(t) + J(t)$. Let the process of reception consist of computing the $K + 1$ integrals

$$x_k = 2T^{-1} \int_{-T/2}^{T/2} [I_k(t) - I_0(t) - J(t)]^2 \, dt, \qquad k = 0, 1, \cdots, K \tag{2-6}$$

and selecting the smallest one (all of the $K + 1$ encodings have been carried to the receiver beforehand). If the value of k corresponding to the smallest integral happens to be 0, as it will be if the noise $J(t)$ is small, no error is made. In any other case the receiver picks out the wrong message.

When the representations (2-2), (2-3), and (2-5) are put in (2-6) and the integrations performed, it is found that

$$x_k = \sum_{n=-N}^{N} (A_n^{(k)} - A_n^{(0)} - B_n)^2, \qquad x_0 = \sum_{n=-N}^{N} B_n^2 \tag{2-7}$$

which have already appeared in equations (1-2) and (1-3). If, as in Section 1, P_k is interpreted as a point in $2N + 1$ — dimensional Euclidean space with coordinates $A_{-N}^{(k)}, \cdots, A_0^{(k)}, \cdots, A_N^{(k)}$ and Q is the point $A_{-N}^{(0)} + B_{-N}, \cdots, A_0^{(0)} + B_0, \ldots, A_N^{(0)} + B_N$, then x_k is the square of the distance between points P_k and Q. Point P_0 corresponds to the signal actually sent, points $P_1, \cdots, P_K$ to the remaining signals, and point Q to the signal plus noise at the receiver. The expected distance between the origin and P_k is $\sigma(2N + 1)^{1/2} = (2W_S)^{1/2}$, that between P_0 and Q is $\nu(2N + 1)^{1/2} = (2W_N)^{1/2}$, and that between the origin and Q is

$$(\sigma^2 + \nu^2)^{1/2}(2N + 1)^{1/2} = (2W_N + 2W_S)^{1/2}$$

No error is made when x_0 is less than every one of $x_1, x_2, \cdots, x_K$, i.e., when none of the points $P_1, \cdots, P_K$ lies within the sphere S of radius $x_0^{1/2}$ centered on Q and passing through P_0. Therefore the probability of obtaining no error when the first encoding scheme is used is equal to the probability denoted by Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$ in the mathematical problem of Section 1.

One might wonder why probability theory has played such a prominent part in the encoding scheme just described. It is used because we do not know the best method of encoding. In fact, it would not be used if we knew how to solve the following problem:* Arrange $K + 1$ points $P_0, \cdots P_K$ on the hyper-surface of the $2N + 1$ — dimensional sphere of radius $(2W_S)^{1/2}$

* C. E. Shannon has commented that although the solution of this problem leads to a good code, it may not be the best possible, i.e., it is not obvious that the code obtained in this way is the same as the one obtained by choosing a set of points so as to minimize the probability of error (calculated from the given set of points and some given W_N) averaged over all $K + 1$ points.

in such a way that the smallest of the $K(K+1)/2$ distances P_kP_ℓ, $k, \ell = 0, 1, \cdots, K$, $k \neq l$, has the largest possible value. This would maximize the difference (as measured by the distance between their representative points) between the two (or more) most similar encoding signals.†

In this paper we have been forced to rely on the randomness of probability theory to secure a more or less uniform scattering of the points $P_0, \cdots, P_K$. In our work they do not lie exactly on a sphere of radius $(2W_S)^{1/2}$ but this causes us no trouble.

3. The Second Encoding Scheme

The second of the two encoding schemes is suggested by one of Shannon's (2) proofs of the fundamental result (1-1). In this scheme the $K+1$ messages are to be sent over a transmission system having a frequency band extending from zero to F cycles per second, and are to be sent during a time interval of nominal length T.

The first few steps in the encoding process are just the same as in the first scheme. N is still given by (2-1) and σ by $(N + 1/2)\sigma^2 = W_S$. After drawing $K+1$ sets of A's, with $2N+1$ in each set, the $K+1$ messages are encoded so that the signal corresponding to the kth message, $k = 0, 1, \cdots, K$, is

$$I_k(t) = (FT)^{1/2} \sum_{n=-N}^{N} A_n^{(k)} \frac{\sin \pi(2Ft - n)}{\pi(2Ft - n)} \tag{3-1}$$

From (3-1), the value of $I_k(t)$ at $t = n/(2F)$ is zero if the integer n exceeds N in absolute value. If the integer n is such that $|n| \leq N$, the corresponding value of $I_k(t)$ is $(FT)^{1/2}A_n^{(k)}$. The energy in the kth signal is obtained by squaring both sides of (3-1) and integrating with respect to t. Thus

$$\int_{-\infty}^{\infty} I_k^2(t)\, dt = 2^{-1}T \sum_{n=-N}^{N} A_n^{(k)2} \tag{3-2}$$

which has the expected value $(N + 1/2)\sigma^2T$. The average power developed when this amount of energy is expended during the nominal signal length T is $(N + 1/2)\sigma^2$ which is equal to W_S, as it should be.

The noise introduced by the transmission system is taken to be

$$J(t) = (FT)^{1/2} \sum_{n=-N}^{N} B_n \frac{\sin \pi(2Ft - n)}{\pi(2Ft - n)} \tag{3-3}$$

† Possibly if $K+1$ discrete unit charges of electricity were allowed to move freely on the sphere, their mutual repulsion would separate them in the required manner. In $2N+1$ dimensions this leads to the problem of minimizing the mutual potential energy

$$\tfrac{1}{2}\Sigma(\overline{P_kP_\ell})^{-2N+1}$$

where $N \geq 1$ and the summation extends over $k, \ell = 0, 1, \ldots K$ with $k \neq \ell$. However, this problem also appears to be difficult.

where the ν universe from which the B's are drawn has, as before, standard deviation ν given by $(N + 1/2)\nu^2 = W_N$. When the signal $I_0(t)$ is sent, the input to the receiver is $I_0(t) + J(t)$ and the process of reception consists of selecting the smallest of the $K + 1$ x_k's

$$x_k = 2T^{-1} \int_{-\infty}^{\infty} [I_k(t) - I_0(t) - J(t)]^2 \, dt \tag{3-4}$$

$$= \sum_{n=-N}^{N} (A_n^{(k)} - A_n^{(o)} - B_n)^2$$

The second expression for x_k is the same as the one given by (2-7) for the first encoding scheme, and the discussion in Section 2 following (2-7) may also be applied to the second encoding scheme. In particular, the probability of obtaining no error in transmitting a signal through noise is the same in both systems of encoding, and is given by the Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$ of the mathematical problem of Section 1.

4. Solution of the Mathematical Problem

We shall simplify the work of solving the mathematical problem stated in Section 1 by taking $\sigma = 1$ and $\nu^2/\sigma^2 = r$. First regard the $4N + 2$ numbers $A_n^{(0)}$, B_n, $n = -N, \cdots, N$ as fixed or given beforehand. Geometrically, this corresponds to having the points P_0 and Q given. Select a typical set of random variables $A_n^{(k)}$, $n = -N, \cdots, N$, $k > 0$ and consider the associated set of variables

$$y_n = A_n^{(k)} - A_n^{(0)} - B_n = A_n^{(k)} + \bar{y}_n. \tag{4-1}$$

y_n is a random variable distributed normally about its average value

$$\bar{y}_n = -A_n^{(0)} - B_n \tag{4-2}$$

with standard deviation $\sigma = 1$. The quantity x_k, defined by (1-3) and representing the square of the distance between P_k and Q, may be written as

$$x_k = \sum_{n=-N}^{N} y_n^2 \tag{4-3}$$

Thus x_k is the sum of the squares of $2N + 1$ independent and normally distributed variates, having the same standard deviation but different average values. The probability density of such a sum is remarkable in that it does not depend upon the $\bar{y}_n$'s individually but only on the sum of their squares which we denote by

$$u = \sum_{n=-N}^{n} \bar{y}_n^2 = \sum_{n=-N}^{n} (A_n^{(0)} + B_n)^2$$

$$= 2T^{-1}\left[\begin{matrix}\text{Energy in sent signal} + \text{Energy} \\ \text{in noise}\end{matrix}\right] \tag{4-4}$$

This behavior follows from the fact that the probability density of P_k has spherical symmetry about the origin (because all the $A_n^{(k)}$'s have the same σ). For the probability that x_k is less than some given value x is the probability that P_k lies within a sphere of radius $x^{1/2}$ centered on Q, and this, because of the symmetry, depends only on x and the distance $u^{1/2}$ of Q from the origin. Accordingly, we write $p(x, u)dx$ for the probability that $x < x_k < x + dx$ when the $\bar{y}_n'$'s (and hence u) are fixed.

The probability density $p(x, u)$ may be obtained from its characteristic function:

$$p(x, u) = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{-izx} [\text{ave. } e^{izx}]\, dz$$

$$\text{ave. } e^{izx} = \text{ave. exp}\left[iz \sum_{n=-N}^{N} y_n^2 \right] \tag{4-5}$$

$$= \prod_{n=-N}^{N} \text{ave. exp } [izy_n^2] = (1 - 2iz)^{-N-1/2} \exp [iuz(1 - 2iz)^{-1}]$$

where we have used (4-3) and, since y_n is distributed normally about $\bar{y}_n$,

$$\text{ave. exp } [izy_n^2] = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{izy_n^2 - (y_n - \bar{y}_n)^2/2}\, dy_n$$

$$= (1 - 2iz)^{-1/2} \exp [\bar{y}_n^2\, iz(1 - 2iz)^{-1}]$$

Hence

$$p(x, u) = (2\pi)^{-1} \int_{-\infty}^{\infty} (1 - 2iz)^{-N-1/2} \exp [izu(1 - 2iz)^{-1} - izx]\, dz \tag{4-6}$$

$$= 2^{-1}(x/u)^{N/2-1/4} I_{N-1/2}[(ux)^{1/2}]\, e^{-(u+x)/2}$$

where it is to be understood that x is never negative. The Bessel function of imaginary argument appears when we change the variable of integration from z to t by means of $1 - 2iz = 2t/x$, and bend the path of integration to the left in the t plane (6). This expression for the probability density of the sum of the squares of a number of normal variates having the same standard deviation but different averages has been given by R. A. Fisher (7).

We are now in a position to solve the following problem which is somewhat simpler than the one stated in Section 1: Given the $2N + 1$ coordinates $A_n^{(0)}$ of the point P_0 and the $2N + 1$ numbers B_n so that the coordinates $A_n^{(0)} + B_n$ of the point Q are given. What is the probability that none of the K points $P_1, P_2, \cdots, P_K$, whose coordinates $A_n^{(k)}$ are drawn at random from a universe distributed normally about zero with standard deviation $\sigma = 1$, be inside the sphere centered on the given point Q and passing through the other given point P_0? In other words, what is the probability that all K of the

independent random variables $x_1, x_2, \cdots, x_K$ will exceed the given value x_0 when u has the value defined by (4-4) together with the given values of the $A_n^{(0)}$'s and B_n's? The variables $x_1, x_2, \cdots, x_K$ have the probability density $p(x, u)$ shown in (4-6) and x_0 is defined by (1-2) and the given values of the B_n's.

The answer to the above problem follows at once when we note that the probability of any one of $x_1, \cdots, x_K$, say x_1 for example, being less than x_0 is

$$P(x_0, u) = \int_0^{x_0} p(x, u)\, dx. \tag{4-7}$$

The probability of x_1 exceeding x_0 is then $1 - P(x_0, u)$ and the probability of all K of $x_1, \cdots, x_K$ exceeding x_0 is

$$[1 - P(x_0, u)]^K \tag{4-8}$$

Instead of being assigned quantities, x_0 and u are actually random variables when we consider the problem of Section 1. Now we take up the problem of finding the probability density of u when x_0 is fixed. Thus, from (4-4), we wish to find the probability density of

$$u = \sum_{n=-N}^{N} (A_n^{(0)} + B_n)^2 \tag{4-9}$$

in which the $2N + 1$ numbers $A_n^{(0)}$ are drawn at random from a universe distributed normally about zero with standard deviation $\sigma = 1$ and the numbers $B_{-N}, \cdots, B_0, \cdots, B_N$ are given. It is seen that u is the sum of the squares of $2N + 1$ normal variates all having the standard deviation $\sigma = 1$. The nth variate, $A_n^{(0)} + B_n$, has the average value B_n. This is just the problem which was encountered at the beginning of this section. Equation (4-9) is of the same form as (4-3) and we have the following correspondence:

Equation (4-3)	*Equation (4-9)*
x_k	u
y_n	$A_n^{(0)} + B_n$
$\bar{y}_n$	B_n
$u = \sum \bar{y}_n^2$	$x_0 = \sum B_n^2$

The probability that u lies in the interval $u, u + du$ when x_0 is given is therefore $p(u, x_0)\, du$ where $p(u, x_0)$ is obtained by putting u for x and x_0 for u in the probability density $p(x, u)$.

Until now x_0 has been fixed. At this stage we regard $B_{-N}, \cdots, B_0, \cdots, B_N$ as random variables drawn from a normal universe of average zero and standard deviation $\nu = \sigma r^{1/2} = r^{1/2}$. If the standard deviation were unity,

the probability density of x_0 could be obtained directly from $p(x, u)$ by letting $u \to 0$ in (4-6). As it is, the x's appearing in the resulting expression must be divided by r to obtain the correct expression. Thus, the probability of finding x_0 between x_0 and $x_0 + dx_0$ is

$$p_0(x_0) = \frac{[x_0/2r]^{N-1/2}}{2r\Gamma(N + 1/2)} \cdot e^{-x_0/2r} \tag{4-10}$$

which is of the χ^2 type frequently encountered in statistical theory.

It follows that the probability of finding u in $(u, u + du)$ and x_0 in $(x_0, x_0 + dx_0)$ at the same time is $p_0(u, x_0)\, du\, dx_0$ where

$$\begin{aligned} p_0(u, x_0) &= p(u, x_0)p_0(x_0) \\ &= \frac{1}{4r\Gamma(N + 1/2)} \left(\frac{ux_0}{4r^2}\right)^{N/2-1/4} I_{N-1/2}[(x_0 u)^{1/2}]\, e^{-[u+x_0(1+1/r)]/2} \end{aligned} \tag{4-11}$$

The replacement of (x, u) in (4-6) by (u, x_0) should be noted.

Now that we have the probability density of u and x_0 we may combine it with the probability (4-8) that all K of $x_1, \cdots, x_K$ exceed x_0 when x_0 and u are fixed. The result is the answer to the problem stated in Section 1:

$$\begin{aligned} &\text{Prob. } (P_1Q, \cdots, P_KQ > P_0Q) \\ &\quad = \int_0^\infty du \int_0^\infty dx_0\, p_0(u, x_0)[1 - P(x_0, u)]^K \end{aligned} \tag{4-12}$$

This result is more complicated than it seems, for $p_0(u, x_0)$ is given by (4-11) and $P(x_0, u)$ is obtained by integrating $p(x, u)$ of (4-6) from $x = 0$ to $x = x_0$ in accordance with (4-7). The remaining portion of the paper is concerned with obtaining an approximation to (4-12) which holds when N and K are very large numbers.

5. Behavior of Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$ as N and K Become Large

In this section we introduce a number of approximations which lead to a manageable expression for Prob. $(P_1Q, \cdots, P_KQ > P_0Q)$ when N and K become large.

Since u and x_0 are sums of independent random variables, namely

$$\begin{aligned} u &= \sum_{n=-N}^{N} (A_n^{(0)} + B_n)^2 \\ x_0 &= \sum_{n=-N}^{N} B_n^2 , \end{aligned} \tag{5-1}$$

the central limit theorem tells us that the probability density $p_0(u, x_0)$ approaches a two-dimensional normal distribution centered on the average

values

$$\begin{aligned} \bar{u} &= \sum_{n=-N}^{N} \text{ave.}\,[A_n^{(0)2} + B_n^2] = (2N+1)(1+r) \\ \bar{x}_0 &= \sum_{n=-N}^{N} \text{ave.}\, B_n^2 = (2N+1)r \end{aligned} \tag{5-2}$$

Here we keep the convention $\sigma = 1$, $\nu^2/\sigma^2 = r$ used in Section 4. The same sort of reasoning as used to establish (5-2) shows that the spread about these average values is given by

$$\begin{aligned} \text{ave. } (u - \bar{u})^2 &= (4N+2)(1+r)^2 \\ \text{ave. } (x_0 - \bar{x}_0)^2 &= (4N+2)r^2 \\ \text{ave. } (u - \bar{u})(x_0 - \bar{x}_0) &= (4N+2)r^2 \end{aligned} \tag{5-3}$$

If the parameters N, K, and r in the integral (4-12) are such that its value is appreciably different from zero, most of the contribution comes from the region around $\bar{u}$ and $\bar{x}_0$ where $p_0(u, x_0)$ is appreciably different from zero. However, instead of taking $\bar{u}$ and $\bar{x}_0$ as reference values, we take the nearby values

$$\begin{aligned} u_2 &= \bar{u} - 2 - 2r = (2N-1)(1+r) = 2q(1+r) \\ x_2 &= \bar{x}_0 - 2r = (2N-1)r = 2qr \end{aligned} \tag{5-4}$$

as these turn out to be better representatives of the center of the distribution. We have introduced the number

$$q = N - 1/2 \tag{5-5}$$

in order to simplify the writing of later equations. We assume $q > 1$.

First, we shall show that

$$\begin{aligned} &\text{Prob. } (P_1 Q, \cdots, P_K Q > P_0 Q) \\ &\quad = \int_{u_2-a}^{u_2+a} du \int_{x_2-b}^{x_2+b} dx_0\, p_0(u, x_0)[1 - P(x_0, u)]^K + R_1 \end{aligned} \tag{5-6}$$

where $a = 2(1+r)(2q \log q)^{1/2}$, $b = 2r(2q \log q)^{1/2}$ and R_1 is of order $1/q$ (denoted by $0(1/q)$), i.e. a constant C and a value q_0 can be found such that $|R_1| < C/q$ when $q > q_0$. From (4-12) it is seen that R_1 is positive and less than

$$\begin{aligned} &\left[\int_0^{u_2-a} du + \int_{u_2+a}^{\infty} du\right] \int_0^{\infty} dx_0\, p_0(u, x_0) \\ &\qquad + \left[\int_0^{x_2-b} dx_0 + \int_{x_2+b}^{\infty} dx_0\right] \int_0^{\infty} du\, p_0(u, x_0) \end{aligned} \tag{5-7}$$

Since $p_0(u, x_0)$ is the joint probability density of u and x_0, the integration with respect to x_0 in the first part of (5-7) yields the probability density of u, and the integration with respect to u in the second part gives the probability density $p_0(x_0)$ (stated in (4-10)) of x_0. Thus (8)

$$\begin{aligned} \int_0^\infty dx_0\, p_0(u, x_0) &= \frac{[u/2(1+r)]^q\, e^{-u/2(1+r)}}{2(1+r)\Gamma(q+1)} \\ \int_0^\infty du\, p_0(u, x_0) &= \frac{[x_0/2r]^q\, e^{-x_0/2r}}{2r\Gamma(q+1)} \end{aligned} \tag{5-8}$$

Setting (5-8) in (5-7) and putting $u = 2(1+r)y$ and $x_0 = 2ry$ in the two parts of (5-7) reduces them to the same form. Thus (5-7) is equal to

$$2 - \frac{2}{\Gamma(q+1)} \int_{q-\ell}^{q+\ell} y^q\, e^{-y}\, dy \tag{5-9}$$

with $\ell = (2q \log q)^{1/2}$. In order to show that (5-9) is $0(1/q)$ we use the expansion

$$-y + q \log y = -q + q \log q - (y-q)^2/(2q) + (y-q)^3/(3q^2) - (y-q)^4 q[q + (y-q)\theta]^{-4}/4$$

where $0 \leqslant \theta \leqslant 1$. Let v represent the sum of the $(y-q)^3$ and $(y-q)^4$ terms, and expand $\exp v$ as $1 + v$ plus a remainder term. The integral of $\exp -(y-q)^2/(2q)$, taken between the limits $q \pm \ell$, can be shown to be of the form $1 - 0(1/q)$ by integrating by parts as in obtaining the asymptotic expansion for the error function. The term in $(y-q)^3$ vanishes upon integration and the remainder terms may be shown to be of $0(1/q)$. In all of this work a square root of q comes in through the fact that

$$1 > (2\pi q)^{1/2} q^q e^{-q}/\Gamma(q+1) > \exp[-1/(12q)] \tag{5-10}$$

We have just shown that the error introduced by restricting the region of integration as indicated by (5-6) introduces an error of order $1/q$ which vanishes as $q \to \infty$. The normal law approximation to $p_0(u, x_0)$ predicted by the central limit theorem holds over this restricted region. However, instead of appealing to the central limit theorem to determine the accuracy of the approximation, we prefer to deal directly with the functions involved.

Consideration of (5-4) and the behavior of $p_0(u, x_0)$ suggests the substitution

$$\begin{aligned} x_0 &= 2r(q+\alpha) \\ u &= 2(1+r)(q+\beta) \end{aligned} \tag{5-11}$$

where α and β are new variables whose absolute values never exceed $(2\ q\ \log\ q)^{1/2}$ in the restricted region of integration of (5-6). From (4-11)

$$p_0(u, x_0)\ du\ dx_0 = \frac{(1 + r)}{\Gamma(q + 1)} \left(\frac{z}{4r^2}\right)^{q/2} I_q(z^{1/2}) e^{-(1+r)(2q+\alpha+\beta)}\ d\alpha\ d\beta \tag{5-12}$$

in which

$$z = ux_0 = 4r(1 + r)(q + \alpha)(q + \beta) \tag{5-13}$$

In Appendix II it is shown that

$$I_q(z^{1/2}) = \frac{q^{q+1/2}\ e^{-q}\ z^{q/2} \exp\ [(q^2 + z)^{1/2} + V]}{\Gamma(q + 1)(q^2 + z)^{1/4}[q + (q^2 + z)^{1/2}]^q} \tag{5-14}$$

where $|\ V\ | < 1/(2q - 1)$ when $q > 1$. Upon using (5-10) and (5-14) the right hand side of (5-12) may be written as

$$d\alpha\ d\beta(2\pi)^{-1/2}(1 + r)(2r)^{-q}(q^2 + z)^{-1/4} \exp\ [-(1 + r)(2q + \alpha + \beta) + f(z) - \log\ \Gamma(q + 1) + 0(1/q)] \tag{5-15}$$

with

$$f(z) = q\ \log\ z - q\ \log\ [q + (q^2 + z)^{1/2}] + (q^2 + z)^{1/2} \tag{5-16}$$

The value z_2 of z corresponding to the central point (u_2, x_2) of $p_0(u, x_0)$ is obtained by putting $\alpha = \beta = 0$ in (5-13):

$$\begin{aligned} z_2 &= 4r(1 + r)q^2 \\ z - z_2 &= 4r(1 + r)[q(\alpha + \beta) + \alpha\beta]. \end{aligned} \tag{5-17}$$

Since we are interested in the form of $p_0(u, x_0)$ in the restricted region of integration of (5-6) we expand $f(z)$ about $z = z_2$ in a Taylor's series plus a remainder term.

$$\begin{aligned} f(z) &= q\ \log\ 2rq + q(1 + 2r) + (z - z_2)/(4rq) \\ &\quad - \frac{(z - z_2)^2}{32r^2q^3(1 + 2r)} + \frac{(z - z_2)^3}{3!} \left[\frac{(\xi_3 + q)^3(3\xi_3 - q)}{8\ z_3^3\ \xi_3^3}\right] \end{aligned} \tag{5-18}$$

In the last term $z_3 = z_2 + (z - z_2)\theta$, $0 \leqslant \theta \leqslant 1$, $\xi_3^2 = q^2 + z_3$. The work of obtaining this expansion is simplified if $(q^2 + z)^{1/2}$ is replaced by ξ in (5-16) before differentiating. For example, by using $2\xi'\xi = 1$, it can be shown that $f'(z)$ is simply $(q + \xi)/(2z)$. When the extreme values of α and β are put in (5-17), it is seen that $z - z_2$ does not exceed $0(q^{3/2}\ \log^{1/2}\ q)$ in the restricted region of integration. In the last term of (5-18) z_3 is $0(q^2)$, ξ_3 is $0(q)$ and consequently the last term itself is $0(q^{-1/2}\ \log^{3/2}\ q)$.

When the expression (5-17) for $(z - z_2)$ is put in (5-18) an expression for $f(z)$ is obtained. This expression, together with

$$\log \Gamma(q + 1) = (q + 1/2) \log q - q + (1/2) \log 2\pi + 0(1/q),$$

enables us to write the argument of the exponential function in (5-15) as $q \log 2r - (1/2) \log 2\pi q - Q(\alpha, \beta) + 0(q^{-1/2} \log^{3/2} q)$ where $Q(\alpha, \beta)$ denotes the quadratic function

$$\begin{aligned} Q(\alpha, \beta) &= [(1 + r)^2(\alpha^2 + \beta^2) - 2r(1 + r)\alpha\beta]D \\ D &= 1/[2q(1 + 2r)] \end{aligned} \tag{5-19}$$

Similar considerations show that

$$(q^2 + z)^{-1/4} = q^{-1/2}(1 + 2r)^{-1/2}[1 + 0(q^{-1/2} \log^{1/2} q)] \tag{5-20}$$

When the above results are gathered together it is found that (5-12) may be written as

$$p_0(u, x_0)\, du\, dx_0 = D_1 \exp [-Q(\alpha, \beta) + 0(q^{-1/2} \log^{3/2} q)]\, d\alpha\, d\beta \tag{5-21}$$

where

$$D_1 = \frac{1 + r}{2\pi q(1 + 2r)^{1/2}} \tag{5-22}$$

Expression (5-21) is valid as long as $|\alpha|$ and $|\beta|$ do not exceed

$$(2q \log q)^{1/2}.$$

Expression (5-21) differs from the one predicted by the central limit theorem (and (5-2) and (5-3)) in that it is not quite centered on the average values $\bar{x}_0$, $\bar{u}$, which correspond to $\alpha = 1$, $\beta = 1$, respectively. Also, q enters in place of $q + 1$. However, these differences amount to $0(q^{1/2} \log^{1/2} q)$ at most, as may be seen by putting $\alpha - 1$ and $\beta - 1$ for α and β in (5-19).

By using relations (5-6) and (5-21), it may be shown that

$$\begin{aligned} &\text{Prob. } (P_1Q, \cdots, P_KQ > P_0Q) \\ &= \int_{-q}^{\infty} d\alpha \int_{-\zeta}^{\infty} d\beta\, D_1\, e^{-Q(\alpha,\beta)}[1 - P(x_0, u)]^K + 0(q^{-1/2} \log^{3/2} q) \end{aligned} \tag{5-23}$$

where it is understood that x_0 and u in $P(x_0, u)$ depend on α and β through (5-11). The term $0(q^{-1/2} \log^{3/2} q)$ in (5-23) represents the sum of three contributions. The first is R_1 in (5-6) which is $0(1/q)$. The second arises from the fact that when the factor $\exp [0(q^{-1/2} \log^{3/2} q)]$ in (5-21) is neglected in integrating (5-21) over $-\ell < \alpha < \ell$, $-\ell < \beta < \ell$, where $\ell = (2q \log q)^{1/2}$, the resulting integral is in error by $0(q^{-1/2} \log^{3/2} q)$. The third is due to the contributions of the integral from the region $|\alpha| > \ell$, $|\beta| > \ell$.

By introducing polar coordinates $\alpha = \rho \cos \theta$, $\beta = \rho \sin \theta$ it can be shown that the region $\rho > \ell$ more than covers the region in question and that

$$Q(\alpha, \beta) \geqslant (1 + r)\rho^2 D \tag{5-24}$$

Upon integrating with respect to ρ and setting in the lower limit ℓ, it is seen that the third contribution is $0(q^{-1/2})$.

We now assume K to be large. Since $0 \leqslant P(x_0, u) \leqslant 1$ we have

$$0 \leqslant e^{-KP} - (1 - P)^K \leqslant KP^2 e^{-KP} < 1/K \tag{5-25}$$

The last inequality follows from $x^2 \exp(-x) < 1$ for $x \geqslant 0$. A proof of the remaining portions will be found in "Modern Analysis" by Whittaker and Watson, Cambridge University Press, Fourth Edition (1927), page 242. When we observe that replacing $[1 - P(x_0, u)]^K$ by $1/K$ in the right hand side of (5-23) gives an integral whose value is less than $1/K$, we see that

$$\text{Prob.}\,(P_1 Q, \cdots, P_K Q > P_0 Q) \tag{5-26}$$

$$= \int_{-q}^{\infty} d\alpha \int_{-q}^{\infty} d\beta\, D_1\, e^{-Q(\alpha,\beta) - KP(x_0, u)} + 0(1/K) + 0(q^{-1/2} \log^{3/2} q)$$

We now take up the problem of expressing the cumulative probability density $P(x_0, u)$ in terms of α and β. When x_0 and u lie in the restricted region of integration shown in (5-6) they are near their average values $\bar{x}_0 = (2N + 1)r$ and $\bar{u} = (2N + 1)(1 + r)$. On the other hand the average value $\bar{x}$ of x and the mean square value σ_x^2 of $(x - \bar{x})^2$ as computed from (4-6), or directly, are $2N + 1 + u$ and $4N + 2 + 4u$, respectively. Thus we see that $\bar{x} - x_0$ is of the same magnitude as $4N$ and becomes much larger than σ_x as $N \to \infty$. The asymptotic development of Appendix I may therefore be used. In Appendix I (equations (A1-27) and (A1-29)) it is shown that when $M(= 2m = 2N + 1)$ is a large number and $1 << (\bar{x} - x_0)\ \sigma_x$

$$P(x_0, u) = (4\pi m b_2)^{-1/2}\ (1 + 0(1/m))\ \exp\,[m\mathrm{F}(v_1)] \tag{5-27}$$

where we have introduced the number $m = N + 1/2 = q + 1$ to save writing $N + 1/2$ or $q + 1$ repeatedly and where

$$\begin{aligned} 2b_2 &= (1 - 1/v_1)^2(1 + 4st)^{1/2} \\ v_1 &= [1 + (1 + 4st)^{1/2}]/2s \\ F(v_1) &= (1 + 4st)^{1/2} - s - t - \log v_1 \\ x_0 &= 2ms = (2N + 1)s, \qquad u = 2mt = (2N + 1)t \end{aligned} \tag{5-28}$$

Comparison of the last line in (5-28) with (5-11) shows that ms and mt are equal to $r(q + \alpha) = r(m + \alpha - 1)$ and

$$(1 + r)(q + \beta) = (1 + r)(m + \beta - 1),$$

respectively. It is convenient to introduce the notation

$$\begin{aligned} \gamma &= \alpha - 1, & \delta &= \beta - 1 \\ s &= r(1 + \gamma/m), & t &= (1 + r)(1 + \delta/m). \end{aligned} \tag{5-29}$$

It is seen that for the restricted region in which $|\alpha|$ and $|\beta|$ are less than $\ell = (2q \log q)^{1/2}$, $|\gamma|$ and $|\delta|$ are at most

$$0(q^{1/2} \log^{1/2} q) = 0(m^{1/2} \log^{1/2} m).$$

Hence s, t, $(1 + 4st)^{1/2}$, v_1 differ at most from r, $1 + r$, $1 + 2r$, $1 + 1/r$, respectively, by terms of order $m^{-1/2} \log^{1/2} m$. Similar considerations show that

$$(4\pi m b_2)^{-1/2} = (2\pi q)^{1/2} D_1[1 + 0(m^{-1/2} \log^{1/2} m)] \tag{5-30}$$

The argument of the exponential function in (5-27) must be expanded in powers of γ and δ. It turns out that when γ and δ lie in the restricted region, powers above the second may be neglected. For the sake of convenience we rewrite (5-13) and introduce z_1:

$$\begin{aligned} z &= x_0 u = 4m^2 st = 4r(1 + r)(m + \gamma)(m + \delta) \\ z_1 &= 4r(1 + r)m^2 \\ z - z_1 &= 4r(1 + r)[m(\gamma + \delta) + \gamma\delta] \end{aligned} \tag{5-31}$$

so that $z - z_1$ is $0(m^{3/2} \log^{1/2} m)$. Then

$$\begin{aligned} (1 + 4st)^{1/2} &= (1 + z/m^2)^{1/2} \\ &= (1 + z_1/m^2)^{1/2} + (z - z_1)(1 + z_1/m^2)^{-1/2}/(2m^2) \\ &\quad - (z - z_1)^2(1 + z_1/m^2)^{-3/2}/(8m^4) + R_2 \end{aligned} \tag{5-32}$$

where R_2 is of the same order as $(z - z_1)^3/m^6$, or $m^{-3/2} \log^{3/2} m$. It follows that

$$\begin{aligned} (1 + 4st)^{1/2} &= 1 + 2r + \frac{2r(1 + r)}{1 + 2r}\left[\frac{\gamma + \delta}{m} + \frac{\gamma\delta}{m^2}\right] \\ &\quad - \frac{2r^2(1 + r)^2}{(1 + 2r)^3}\frac{(\gamma + \delta)^2}{m^2} + 0(m^{-3/2} \log^{3/2} m) \\ v_1 &= \frac{(1 + r)}{r(1 + \gamma/m)}\left\{1 + \frac{r}{1 + 2r}\left[\frac{\gamma + \delta}{m} + \frac{\gamma\delta}{m^2}\right]\right. \\ &\quad \left. - \frac{r^2(1 + r)(\gamma + \delta)^2}{m^2(1 + 2r)^2} + 0(m^{-3/2} \log^{3/2} m)\right\}. \end{aligned} \tag{5-33}$$

Combining these and a similar expression for $\log v_1$ leads to

$$\begin{aligned} mF(v_1) &= -m \log (1 + 1/r) + \gamma - \delta \\ &\quad -[(1 + r)\gamma - r\delta]^2/[2m(1 + 2r)] + 0(m^{-1/2} \log^{3/2} m) \\ &= -(q + 1) \log (1 + 1/r) + \alpha - \beta - [(1 + r)\alpha - r\beta]^2 D \\ &\qquad + 0(q^{-1/2} \log^{3/2} q) \end{aligned} \tag{5-34}$$

Substitution of (5-30) and (5-34) in (5-27) gives the result we seek:

$$\begin{aligned} P(x_0, u) &= (1 + 1/r)^{-q-1}(2\pi q)^{1/2} D_1 \\ &\qquad \exp (\alpha - \beta - [(1 + r)\alpha - r\beta]^2 D + 0(q^{-1/2} \log^{3/2} q)) \end{aligned} \tag{5-35}$$

Since $P(x_0, u)$ occurs only in the product $KP(x_0, u)$ in (5-26) we set, in view of (5-35),

$$KP(x_0, u) = A\lambda(\alpha, \beta) \exp S(\alpha, \beta) \tag{5-36}$$

where $\lambda(\alpha, \beta)$ stands for the terms denoted by $\exp [0(q^{-1/2} \log^{3/2} q)]$ in (5-35) and

$$\begin{aligned} A &= K(1 + 1/r)^{-q-1}(2\pi q)^{1/2} D_1 \\ S(\alpha, \beta) &= \alpha - \beta - [(1 + r)\alpha - r\beta]^2 D \end{aligned} \tag{5-37}$$

As long as $|\alpha| < \ell$ and $|\beta| < \ell$, $\lambda(\alpha, \beta)$ is nearly unity and we write

$$\begin{aligned} &\lambda_1 < \lambda(\alpha, \beta) < \lambda_2 \\ &\lambda_1 = 1 - \epsilon, \lambda_2 = 1 + \epsilon, \epsilon = Cq^{-1/2} \log^{3/2} q \end{aligned} \tag{5-38}$$

where C is a positive constant large enough to make ϵ dominate the terms of order $q^{-1/2} \log^{3/2} q$ in (5-35). q is supposed to be so large that ϵ is very small in comparison with unity.

Setting (5-36) in (5-26) gives

$$\text{Prob. } (P_1Q, \cdots, P_\kappa Q > P_0Q) = I + 0(1/K) + 0(q^{-1/2} \log^{3/2} q) \tag{5-39}$$

where the contribution of the region outside $|\alpha| < \ell$, $|\beta| < \ell$ has been returned to the terms denoted by $0(q^{-1/2} \log^{3/2} q)$ (we could have stayed in the region $|\alpha| < \ell$, $|\beta| < \ell$ from (5-23) onward, but didn't do so because we wanted to show that the results coming from (5-25) were not restricted to this region) and

$$I = \int_{-\ell}^{\ell} d\alpha \int_{-\ell}^{\ell} d\beta \, D_1 \exp [-Q(\alpha, \beta) - A\lambda(\alpha, \beta) e^{S(\alpha, \beta)}] \tag{5-40}$$

Let $L(\lambda)$ denote the integral obtained by replacing the function $\lambda(\alpha, \beta)$ in I by the positive constant λ (which we shall take to be either λ_1 or λ_2 defined

by (5-38)). Then, since $A \exp S(\alpha, \beta)$ is positive, it follows from (5-40) that

$$L(\lambda_1) > I > L(\lambda_2) \tag{5-41}$$

Also since $\exp [-A\lambda \exp S(\alpha, \beta)]$ lies between 0 and 1 for all real values of α and β it may be shown from (5-24) that $L(\lambda)$ is equal to $J(\lambda) + 0(q^{-1/2})$ where

$$J(\lambda) = \int_{-\infty}^{\infty} d\alpha \int_{-\infty}^{\infty} d\beta \, D_1 \exp [- Q(\alpha, \beta) - A\lambda e^{S(\alpha,\beta)}] \tag{5-42}$$

Here λ is a constant and $Q(\alpha, \beta)$, A, $S(\alpha, \beta)$ are defined by (5-19) and (5-37). From (5-39) and (5-41) we obtain

$$\text{Prob. } (P_1 Q, \cdots, P_K Q > P_0 Q) = J(1) + \theta[J(\lambda_1) - J(1)] \tag{5-43}$$
$$+ (1 - \theta)[J(\lambda_2) - J(1)] + 0(1/K) + 0(q^{-1/2} \log^{3/2} q)$$

where $0 < \theta < 1$. It will be shown later that $J(\lambda_1)$ and $J(\lambda_2)$ differ from $J(1)$ by terms which are certainly not larger than $0(q^{-1/2})$.

The problem now is to evaluate the integral (5-42) for $J(\lambda)$. It turns out that $\exp [-A\lambda \exp S(\alpha, \beta)]$ acts somewhat like a discontinuous factor which is unity when $S(\alpha, \beta) + \log A\lambda$ is negative and zero when it is positive. In order to investigate this behavior we make the change of variable

$$\begin{aligned} \alpha - \beta &= w \qquad \alpha = y - rw \\ (1 + r)\alpha - r\beta &= y \qquad \beta = y - (1 + r)w \\ d\alpha \, d\beta &= dw \, dy \end{aligned} \tag{5-44}$$

From (5-19), (5-37), and (5-42)

$$\begin{aligned} Q(\alpha, \beta) &= [y^2 + (1 + 2r)\beta^2]D = y^2 D + \beta^2/2q \\ S(\alpha, \beta) &= w - y^2 D \\ J(\lambda) &= \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dw \, D_1 \exp [- y^2 D - \beta^2/2q - A\lambda e^{w - y^2 D}] \end{aligned} \tag{5-45}$$

Here and in the following work β is to be regarded as a function of w and y.

Split the interval of integration with respect to w into the two subintervals $(-\infty, w_0)$ and (w_0, ∞) where

$$w_0 = y^2 D - \log A\lambda \tag{5-46}$$

and y is temporarily regarded as constant. In the first interval

$$\begin{aligned} &\int_{-\infty}^{w_0} \exp [- \beta^2/2q - e^{w - w_0}] \, dw \\ &\quad = \int_{-\infty}^{w_0} e^{-\beta^2/2q} \, dw - \int_{-\infty}^{w_0} (1 - \exp [- e^{w - w_0}]) e^{-\beta^2/2q} \, dw \end{aligned} \tag{5-47}$$

Splitting the interval of integration $(-\infty, w_0)$ into $(-\infty, -\log A\lambda)$ and $(-\log A\lambda, w_0)$ in the first integral on the right of (5-47) shows that its contribution to $J(\lambda)$ is

$$D_1 \int_{-\infty}^{\infty} dy \int_{-\infty}^{-\log A\lambda} dw\, e^{-y^2 D - \beta^2/2q} + D_1 \int_{-\infty}^{\infty} dy \int_{-\log A\lambda}^{w_0} dw\, e^{-y^2 D - \beta^2/2q} \tag{5-48}$$

Integrating with respect to y, after inverting the order of integration, shows that the value of the first integral is

$$\pi^{-1/2} \int_{-\infty}^{B} e^{-t^2}\, dt = (1 + \operatorname{erf} B)/2 \tag{5-49}$$

where, from (5-37) and the definition (5-22) of D_1,

$$\begin{aligned} B &= -\tfrac{1}{2}(1 + r)^{1/2} q^{-1/2} \log A\lambda \\ &= -\tfrac{1}{2}(1 + r)^{1/2} q^{-1/2} \log \frac{\lambda K r (1 + 1/r)^{-q}}{[2\pi q(1 + 2r)]^{1/2}} \end{aligned} \tag{5-50}$$

That the value of $J(\lambda)$ differs from (5-49) by $0(q^{-1/2})$ may be seen as follows. Since $0 < \exp[-\beta^2/2q] < 1$, the integral over (w_0, ∞) (mentioned just above (5-46) and obtained by taking the limits of integration to be w_0 and ∞ in the left side of (5-47)) is positive and less than

$$\int_{w_0}^{\infty} \exp[-e^{w-w_0}]\, dw = \int_1^{\infty} e^{-x}\, dx/x = .219\ldots \tag{5-51}$$

Likewise, the second integral on the right side of (5-47) is less than

$$\int_{-\infty}^{w_0} (1 - \exp[-e^{w-w_0}])\, dw = \int_0^1 (1 - e^{-x})\, dx/x = .796\ldots \tag{5-52}$$

Therefore the contribution of the first integral on the right of (5-47) differs from $J(\lambda)$ by a quantity less than

$$\int_{-\infty}^{\infty} D_1 e^{-y^2 D}(.219 + .796)\, dy = 0(q^{-1/2})$$

in absolute value. The contribution of the first integral on the right of (5-47) differs from (5-49) by the second integral in (5-48) which is $0(q^{-1/2})$ because it is less than

$$\int_{-\infty}^{\infty} D_1(y^2 D) e^{-y^2 D}\, dy$$

The factor $(y^2 D)$ arises from $w_0 - (-\log A\lambda)$ when the mean value theorem is applied to the integral in w. Hence $J(\lambda)$ differs from (5-49) by $0(q^{-1/2})$.

Although (5-49) is a sufficiently accurate expression of $J(\lambda)$ for our pur-

poses, it seems worthwhile to set down approximate expressions for the terms which have been dismissed as $0(q^{-1/2})$. From the above work,

$$J(\lambda) = (1 + \operatorname{erf} B)/2 + D_1 \int_{-\infty}^{\infty} dy\, e^{-y^2 D} \left\{ \int_{w_0}^{\infty} e^{-\beta^2/2q} \exp[-e^{w-w_0}]\, dw \right.$$

$$- \int_{-\infty}^{w_0} e^{-\beta^2/2q} (1 - \exp[-e^{w-w_0}])\, dw$$

$$\left. + \int_{-\log A\lambda}^{w_0} e^{-\beta^2/2q}\, dw \right\} \tag{5-53}$$

$$\approx (1 + \operatorname{erf} B)/2 + D_1 \int_{-\infty}^{\infty} dy\, e^{-y^2 D} \{-.577.. + y^2 D\} e^{-\beta_1^2/2q}$$

$$= (1 + \operatorname{erf} B)/2 + \left(\frac{1+r}{4\pi q}\right)^{1/2} [-.577 \cdots +$$

$$4^{-1}(1 + r)^{-1}\{1 + (2 + 4r)B^2\}] e^{-B^2}$$

where $\beta_1 = y + (1 + r) \log A\lambda$ and we have made use of the fact that $\beta^2/2q$ changes relatively slowly in comparison with w when q is large.

Since $J(\lambda)$ differs from $(1 + \operatorname{erf} B)/2$ by $0(q^{-1/2})$, and since the three B's for λ equal to λ_1, 1, and λ_2 differ by not more than $0(q^{-1/2} \log (\lambda_2/\lambda_1)) = 0(q^{-1} \log^{3/2} q)$, from (5-50) and (5-38), it follows that the terms involving $J(\lambda_1)$ and $J(\lambda_2)$ in (5-43) may be included in the term $0(q^{-1/2} \log^{3/2} q)$. In using our result it is more convenient to deal with N and $K + 1$ instead of $q = N - 1/2$ and K. Hence instead of B we deal with H defined by

$$H = -\frac{1}{2} \frac{(1+r)^{1/2}}{(q + 1/2)^{1/2}} \log \frac{(K+1)(1 + 1/r)^{-q-1}(1+r)}{[2\pi(q + 1/2)(1 + 2r)]^{1/2}}. \tag{5-54}$$

The difference $B - H$, with $\lambda = 1$ and H finite, may be shown to be (with considerable margin) $0(1/K) + 0(q^{-1/2})$. From (5-43), as amended by the first sentence in this paragraph, it follows that

$$\text{Prob.}\,(P_1 Q, \cdots, P_K Q > P_0 Q) = (1 + \operatorname{erf} H)/2 + 0(1/K) + 0(q^{-1/2} \log^{3/2} q) \tag{1-4}$$

where the difference between erf B and erf H has been absorbed by the "order of" terms. When $q + 1/2$ is replaced by N in (5-54) the result is expression (1-5) for H.

APPENDIX I

Cumulative Distribution Function for a Sum of Squares of Normal Variates

Let x be a random variable defined by

$$x = \sum_{n=1}^{M} y_n^2 \tag{A1-1}$$

where y_n is a random variable distributed normally about its average value $\bar{y}_n$ with unit standard deviation. In writing (A1 − 1) we have been guided by (4-3), where $M = 2N + 1$, but here we shall let M be any positive integer. In much of the following work $M/2$ occurs and for convenience we put

$$m = M/2 \tag{A1-2}$$

From the work of Section 4 it follows that the probability density $p(x, u)$ of x is given by Fisher's expression

$$p(x, u) = 2^{-1}(x/u)^{m/2-1/2} I_{m-1}[(ux)^{1/2}]e^{-(u+x)/2} \tag{A1-3}$$

where u is the constant

$$u = \sum_{n=1}^{M} \bar{y}_n^2 \tag{A1-4}$$

Here we are interested in the cumulative distribution function, i.e., the probability that x is less than some given value x_0,

$$P(x_0, u) = \int_0^{x_0} p(x, u)\, dx \tag{A1-5}$$

as M becomes large. In this case the central limit theorem tells us that $p(x, u)$ approaches a normal law with average $\bar{x} = M + u$ and variance = ave. $(x - \bar{x})^2 = 2M + 4u$. The function $P(x_0, u)$ has been studied by J. I. Marcum in some unpublished work, and by P. K. Bose(9). In particular, Marcum has used the Gram-Charlier series to obtain values for $P(x_0, u)$ in the vicinity of $\bar{x}$ for large values of M. However, since I have not been able to find any previous work covering the case of interest here, namely values of $P(x_0, u)$ when x_0 is appreciably less than $\bar{x}$, a separate investigation is necessary and will be given here.

Integrating the general expression (4-5) with respect to x between $-X$ and x_0, letting $X \to \infty$, and discarding the portions of the integrand which oscillate with infinite rapidity gives

$$P(x_0, u) = -\frac{1}{2\pi i}\int_{-\infty,\text{ above }0}^{\infty} z^{-1}e^{-izx_0}[\text{ave. } e^{izx}]\,dz$$
$$= 1 - \frac{1}{2\pi i}\int_{-\infty,\text{ below }0}^{\infty} z^{-1}e^{-izx_0}[\text{ave. } e^{izx}]\,dz \tag{A1-6}$$

where the subscripts "above 0" and "below 0" indicate that the path of integration is indented so as to pass above or below, respectively, the pole at $z = 0$. The value of ave. exp (izx) may be obtained by setting $N + 1/2 = m$ in (4-5). The new notation

$$x_0 = Ms = 2ms, \qquad u = 2mt, \qquad 2z = \zeta \tag{A1-7}$$

enables us to write

$$P(x_0, u) = -\frac{1}{2\pi i}\int_{-\infty,\text{ above }0}^{\infty} \zeta^{-1}\exp\, m[-is\zeta - \log(1 - i\zeta) - t + t(1 - i\zeta)^{-1}]\,d\zeta. \tag{A1-8}$$

The further change of variable

$$1 - i\zeta = v \tag{A1-9}$$

carries (A1-8) into

$$P(x_0, u) = \frac{1}{2\pi i}\int_K (1 - v)^{-1}\exp[mF(v)]\,dv \tag{A1-10}$$

where the path of integration K is the straight line in the complex v plane running from $1 + i\infty$ to $1 - i\infty$ with an indentation to the right of $v = 1$, and

$$F(v) = sv - \log v + t/v - s - t. \tag{A1-11}$$

The K used here should not be confused with the K denoting the number of messages in the body of the paper. We have run out of suitable symbols.

An asymptotic expression for (A1-10) will now be obtained by the method of "steepest descents." The saddle points are obtained by setting the derivative

$$F'(v) = s - 1/v - t/v^2 \tag{A1-12}$$

to zero and are at

$$v_1 = [1 + (1 + 4st)^{1/2}]/2s$$
$$v_2 = [1 - (1 + 4st)^{1/2}]/2s \tag{A1-13}$$

As x_0 and s increase from 0 to ∞, u and t of course being fixed, we have the following behavior:

$$\begin{array}{lll} x_0 = 0 & \bar{x} & \infty \\ s = 0 & 1 + t & \infty \\ v_1 = \infty & 1 & 0 \\ v_2 = -t & -t/(1+t) & 0 \end{array} \tag{A1-14}$$

It is seen that $v_1 \geqslant 0$ and $v_2 \leqslant 0$.

Putting aside for the moment the factor $(1 - v)^{-1}$ in (A1-10), the path of steepest descent through the saddle point v_1 is one of the two curves specified by equating the imaginary part of $F(v)$ to zero. Introducing polar coordinates gives

$$\begin{aligned} v &= \rho e^{i\theta} \\ \text{Real } F(v) &= (s\rho + t/\rho)\cos\theta - \log\rho - s - t \\ \text{Imag. } F(v) &= (s\rho - t/\rho)\sin\theta - \theta \end{aligned} \tag{A1-15}$$

At v_1, $\theta = 0$, $\rho = v_1$. Imag. $F(v_1) = 0$ and, from (A1-12),

$$\begin{aligned} \text{Real } F(v_1) &= (2sv_1 - 1) - \log v_1 - s - t \\ &= (1 + 4st)^{1/2} - \log v_1 - s - t \end{aligned} \tag{A1-16}$$

The path of steepest descent through v_1 may be obtained in polar form by solving

$$(s\rho - t/\rho) = \theta/\sin\theta \tag{A1-17}$$

for ρ as a function of θ. Setting $\varphi = \theta \csc\theta$ and taking the positive value of ρ leads to

$$\rho = [\varphi + (\varphi^2 + 4st)^{1/2}]/2s \tag{A1-18}$$

As θ increases from 0 to π, φ increases from 1 to ∞, and ρ starts from v_1 (as it should) and ends at ∞. Thus, the path of steepest descent through v_1 comes in from $v = -\infty + i\pi/s$ (when θ is nearly π, $\rho \approx \varphi/s$, $\varphi \approx \pi/(\pi - \theta)$ and $\rho(\pi - \theta) \approx \pi/s$), crosses the positive imaginary v axis and bends down to cut the real positive v axis (at right angles) at v_1, and then goes out to $v = -\infty - i\pi/s$ along a similar path in the lower part of the plane. It thus avoids the branch cut (which we take to run from $-\infty$ to 0) in the v plane necessitated by the term $\log v$ in $F(v)$. Since m and s are positive the path of integration K in (A1-10) may be made to coincide with the path of steepest descent when $v_1 > 1$. This corresponds to the case in which $x_0 < \bar{x}$ as (A1-14)

shows. When $0 < v_1 < 1$, i.e., $\infty > x_0 > \tilde{x}$, the two paths may still be made to coincide but it is necessary to add the contribution of the pole at $v = 1$ as K is pulled over it. This is equivalent to passing from the first to the second of equations (A1-6). The path $\theta = 0$ which makes Imag. $F(v)$ of (A1-15) zero turns out to be the curve of "steepest ascent" and hence need not be considered. As (A1-13) shows, the saddle point v_2 does not enter into our considerations because it lies on the negative real v axis and the path of integration K in (A1-10) cannot be made to pass through it without trouble from the singularity of $F(v)$ at $v = 0$.

We now suppose $x_0 < \tilde{x}$ so that s and t are such as to make $v_1 > 1$. In order to remove the factor $(1 - v)$ from the denominator of the integrand in (A1-10), we change the variable of integration from v to w:

$$v - 1 = e^w, \qquad (1 - v)^{-1}dv = -dw$$
$$P(x_0, u) = -\frac{1}{2\pi i}\int_L \exp\,[mF(1 + e^w)]\,dw \tag{A1-19}$$

As v comes in along the path of steepest descent, the path of integration L for w comes in from $w = \infty + i\pi$ and dips down towards the real w axis as arg v decreases from π. L crosses the real w axis perpendicularly at the point

$$w_1 = \log\,(v_1 - 1) \tag{A1-20}$$

and then runs out to $w = \infty - i\pi$ along a curve which tends to become parallel to the real w axis. w_1 may be either positive or negative. When x_0 is almost as large as $\tilde{x}$, w_1 is large and negative.

Since $F(v)$ is real along the path of steepest descent, $F(1 + e^w)$ is real along L. This real value is $-\infty$ at the ends of L and attains its maximum value $F(v_1)$, given by (A1-16), at $w = w_1$. w_1 is a saddle point in the complex w plane because

$$\frac{d}{dw}F(1 + e^w) = F'(1 + e^w)e^w = F'(v)e^w \tag{A1-21}$$

vanishes at $w = w_1$.

Instead of $F(1 + e^w)$ itself we shall be concerned with

$$\tau = F(1 + e^{w_1}) - F(1 + e^w) \tag{A1-22}$$

so that (A1-19) may be written as

$$P(x_0, u) = -\frac{\exp\,[mF(1 + e^{w_1})]}{2\pi i}\int_L e^{-m\tau}\,dw. \tag{A1-23}$$

The variable τ is real on the path of integration L, is zero at w_1, and increases to $+\infty$ as we follow L out to $w = \infty \pm i\pi$. It is convenient to split

K into two parts (10). The first part connects $\infty + i\pi$ to w_1 and the second part connects w_1 to $\infty - i\pi$. The values of w on these two parts will be denoted by w_I and w_{II}, respectively. Corresponding to each value of τ there is a value w_I and a value w_{II} (in fact it turns out that w_{II} is the conjugate complex of w_I). Changing the variable of integration in (A1-23) from w to τ, and remembering that K starts at $\infty + i\pi$, gives

$$P(x_0, u) = \frac{\exp [mF(1 + e^{w_1})]}{2\pi i} \int_0^\infty e^{-m\tau} \left[\frac{d}{d\tau} w_I - \frac{d}{d\tau} w_{II} \right] d\tau \quad \text{(A1-24)}$$

Since m is large, most of the contribution to the value of the integral comes from around $\tau = 0$ or $w = w_1$. In order to obtain an expression for the integrand in this region we note that, because $F'(v_1) = 0$, the Taylor series for (A1-22) is of the form

$$\tau = -b_2(w - w_1)^2 - b_3(w - w_1)^3 - b_4(w - w_1)^4 - \cdots \quad \text{(A1-25)}$$

The circle of convergence of this series is centered on w_1 and extends out to $w = \pm i\pi$, these points being the nearest singularities of $F(1 + e^w)$ as may be seen by setting $v = 1 + e^w$ in (A1-11) and observing that the singularities of $\log v - t/v$ in the finite portion of the w plane occur at odd multiples of $\pm i\pi$. We imagine the branch cuts associated with $\log v$ to run out to the right from these points along lines parallel to the real w axis. Since (A1-25) has a non-zero radius of convergence, the same is true of the two series obtained from it by inversion, namely

$$\begin{aligned} w_I - w_1 = {} & ib_2^{-1/2}\tau^{1/2} + b_3\tau/2b_2^2 \\ & + i[b_2^{-2}b_4 - 5b_2^{-3}b_3^2/4]\tau^{3/2}/2b_2^{1/2} + \cdots \end{aligned} \quad \text{(A1-26)}$$

and the series for $w_{II} - w_1$ obtained from (A1-26) by changing the sign of i. Differentiation of these two series gives a series for $d(w_I - w_{II})/d\tau$ which also converges for sufficiently small $|\tau|$ (putting aside the term in $\tau^{-1/2}$), and which, when put in (A1-24), leads to

$$P(x_0, u) \sim \frac{e^{mF(v_1)}}{(4\pi m b_2)^{1/2}} \left\{ 1 + \frac{3}{4m} [b_2^{-2}b_4 - 5b_2^{-3}b_3^2/4] + \cdots \right\} \quad \text{(A1-27)}$$

That this is an asymptotic expansion holding for large values of m follows from a lemma given by Watson (11). The conditions of the lemma hold since we have already shown that the series for $d(w_I - w_{II})/d\tau$ converges for $|\tau|$ small enough. Furthermore, $d(w_I - w_{II})/d\tau$ is bounded for $a \leqslant \tau$ where τ is real and $0 < a \leqslant$ the radius of convergence of (A1-26). This follows the fact that

$$\frac{dw}{d\tau} = \left[\frac{d\tau}{dw} \right]^{-1} = [-F'(1 + e^w)e^w]^{-1}$$

is bounded except near $w = w_1$ (i.e., $\tau = 0$) and, indeed, decreases to zero like $-e^{-w}/s$ as $w \to \infty \pm i\pi$ (i.e., $\tau \to \infty$).

The values of b_2, b_3, b_4 obtained by expanding (A1-22) and comparing the result with (A1-25) are

$$\begin{aligned} b_2 &= F''(v_1)e^{2w_1}/2 \\ b_3 &= [F'''(v_1)e^{3w_1} + 3F''(v_1)e^{2w_1}]/6 \\ b_4 &= [F''''(v_1)e^{4w_1} + 6F'''(v_1)e^{3w_1} + 7F''(v_1)e^{2w_1}]/24 \end{aligned} \tag{A1-28}$$

$$F''(v) = v^{-2} + 2tv^{-3}, \quad F'''(v) = -2v^{-3} - 6tv^{-4}, \quad F''''(v) = 6v^{-4} + 24tv^{-5}$$

Our asymptotic expression for $P(x_0, u)$, when $x_0 < \bar{x}$, is given by (A1-28) and (A1-27). Only the leading term of (A1-27) is used in the paper. Sometimes the following expressions are more convenient than the ones which have already been given.

$$\begin{aligned} b_2 &= v_1^{-3}(v_1 + 2t)e^{2w_1}/2 = v_1^{-3}(v_1 + 2t)(v_1 - 1)^2/2 \\ &= (1 - 1/v_1)^2(1 + 4st)^{1/2}/2 \\ F(v_1) &= (1 + 4st)^{1/2} - s - t - \log v_1. \end{aligned} \tag{A1-29}$$

In all of these formulas v_1 is given in terms of s and t by (A1-13) and s and t in terms of x_0 and u by (A1-7).

When $x_0 > \bar{x}$, the saddle point v_1 lies between 0 and 1 in the v plane. As v follows the path of steepest descent (discussed just below equation (A1-18)) arg $(v - 1)$ now stays close to π. From (A1-19) Imag. w stays close to π on the new path of steepest descent in the w plane, and the saddle point w_1 now lies on the negative real portion of the line Imag. $w = \pi$. The new path starts at $w = \infty + i\pi$, swings down a little as it comes in, swerves up to pass through w_1 and then goes out to $w = \infty + i\pi$ above the branch cut joining $w = i\pi$ to $w = \infty + i\pi$. The analysis goes along much as for $v_1 > 1$ except that instead of being 0 the imaginary part of w_1 is $i\pi$. This causes the terms in b_3 and b_4 containing exp $(3w_1)$ to change sign. The numerical values of b_2 and $F(v_1)$ are computed by the formulas (A1-29) as before. The fact that b_2 contains the factor exp $(i2\pi)$ shows up only in changing the sign of $b_2^{1/2}$ to give the minus sign in the leading term:

$$P(x_0, u) \sim 1 - (4\pi m|\, b_2\,|)^{-1/2} \exp\,[mF(v_1)]$$

which holds for $x_0 > \bar{x}$. The one arises from the pole at $v = 1$ and is the same as the one in the second of equations (A1-6).

In order to see how (A1-27) breaks down near $x_0 = \bar{x}$, we set $x_0 - \bar{x} = 2m(s - 1 - t) = -2m\epsilon$ or $s = 1 + t - \epsilon$ where ϵ is a small positive number

Using $\sigma_x^2 \equiv$ ave. $(x - \bar{x})^2 = 4(m + u) = 4m(1 + 2t)$ it is found that

$$v_1 = 1 + \epsilon/(1 + 2t) = 1 - 2(x_0 - \bar{x})\sigma_x^2$$

$$mF(v_1) = -m\epsilon^2/(2 + 4t) = -(x_0 - \bar{x})^2/2\sigma_x^2$$

$$2mb_2 = m(v_1 - 1)^2(1 + 2t) = (x_0 - \bar{x})^2/\sigma_x^2$$

and that, since $w_1 \to -\infty$, $b_3 \to b_2$ and $b_4 \to 7b_2/12$. When these values are put in (A1-27) the leading term becomes

$$P(x_0, u) \sim (2\pi)^{-1/2}(\sigma_x/z) \exp [-z^2/2\sigma_x^2]$$

and the term within the braces in (A1-27) reduces to $1 - \sigma_x^2/z^2$ where $z = \bar{x} - x_0 > 0$. Since the asymptotic expansion is useful only in the region where the second term within the braces is small in comparison with the first term, which is unity, $\bar{x} - x_0$ must be several times as large as σ_x before we can use (A1-27). It will be noticed that the above expression for $P(x_0, u)$ is closely related to the asymptotic expansion of the error function.

APPENDIX II

An Approximation for $I_N(x)$

When z in the Bessel function $J_q(qz)$ is imaginary a formula given by Meissel (12) becomes

$$I_q(qy) = \frac{(qy)^q \exp(qw + V)}{\epsilon^q \Gamma(q + 1) w^{1/2}(1 + w)^q} \tag{A2-1}$$

where $w = (1 + y^2)^{1/2}$ and V is a function of y and q which, when q is large, has the formal expansion

$$V = \frac{1}{24q}\left\{2 - \frac{2 - 3y^2}{w^3}\right\} + \frac{y^4 - 4y^2}{16q^2w^6} - \frac{1}{5760q^3}\left\{16 - \frac{16 + 1512y^2 - 3654y^4 + 375y^6}{w^9}\right\} + \cdots \tag{A2-2}$$

Here we shall show that for $y \geqslant 0$ and $q > 1$

$$|V| < 1/(2q - 1) \tag{A2-3}$$

Consideration of (A2-2) and also of the method used to establish (A2-3) indicates that the inequality is very rough. It doubtlessly can be greatly improved (but not beyond the $1/(12q)$ obtained by letting y and $q \to \infty$ in (A2-2)). Incidentally, it may be shown that the constant terms which remain in (A2-2) when $y = \infty$ are associated with the asymptotic expansion of $\log \Gamma(q + 1)$.

When (A2-1) is substituted in Bessel's differential equation, which we write as

$$y^2 \frac{d^2}{dy^2} I_q(qy) + y \frac{d}{dy} I_q(qy) - q^2(1 + y^2) I_q(qy) = 0,$$

we obtain a differential equation for V:

$$V'' = (4 - y^2)w^{-4}/4 - (2qw + w^{-2})y^{-1}V' - V'^2 \qquad \text{(A2-4)}$$

Here the primes denote differentiation with respect to y. The constants of integration associated with (A2-4) are to be chosen so that

$$V \to y^2/(4q + 4) \text{ as } y \to 0. \qquad \text{(A2-5)}$$

This condition is obtained by comparing the limiting form of (A2-1), in which $w \to 1 + y^2/2$, with

$$I_q(qy) \to \frac{(qy/2)^q}{\Gamma(q + 1)} \left[1 + \frac{(qy/2)^2}{q + 1}\right] \to \frac{(qy/2)^q}{\Gamma(q + 1)} \exp\left[\frac{q^2y^2}{4(q + 1)}\right]$$

Condition (A2-5) completely determines V since substitution of the assumed solution

$$V = 4^{-1}(q + 1)^{-1}y^2 + c_1y^4 + c_2y^6 + \cdots$$

in (A2-4) leads to relations which determine $c_1, c_2, \cdots$ successively.

Let $V' = v$. Then (A2-4) becomes

$$v' = c - 2bv - v^2 \qquad \text{(A2-6)}$$

where c and b are known functions of y defined by

$$c = (4 - y^2)w^{-4}/4, \qquad b = (qw + w^{-2}/2)y^{-1} \qquad \text{(A2-7)}$$

From (A2-5), $v \to y/(2q + 2)$ as $y \to 0$ and therefore

$$V = \int_0^y v \, dy \qquad \text{(A2-8)}$$

We first show that $|v| < 1/(2q - 1)$ when $q > 1$. The (y, v) plane may be divided into regions according to the sign of v'. The equations of the dividing lines between these regions are obtained by setting $v' = 0$ in (A2-6). Thus, for a given value of y, v' is positive if $v_2 < v < v_1$ and negative if $v > v_1$ or $v < v_2$ where

$$\begin{aligned} v_1 &= -b + (b^2 + c)^{1/2} = c/[b + (b^2 + c)^{1/2}] \\ v_2 &= -b + (b^2 + c)^{1/2} \end{aligned} \qquad \text{(A2-9)}$$

When $y > 0$ we have $b \geqslant q$. A plot of c versus y shows that $|c| \leqslant 1$. Hence,

when $q > 1$,

$$b^2 + c \geqslant q^2 - 1 > (q-1)^2$$
$$|v_1| < 1/(2q-1) \tag{A2-10}$$
$$v_2 < -2q + 1$$

The curve obtained by plotting v_1 as a function of y plays an important role because, as we shall show, the maxima and minima of the curve for v lie on it. Therefore, the maximum value of $|v|$ cannot exceed the maximum value of $|v_1|$. The maxima and minima must lie on either the v_1 or the v_2 curve since v' vanishes only on these curves. In order to show that it is the v_1 curve we note from (A2-9) that, near $y = 0$, v_1 behaves like $y/(2q+1)$. Consequently both the v_1 and v curves start from $v = 0$ at $y = 0$ but for a while v_1 lies above v which behaves like $y/(2q+2)$. Here v lies in a $v' > 0$ region and continues to increase until it intersects v_1 (as it must do before y reaches 2 because $v_1 = 0$ at $y = 2$) at which point $v' = 0$, $v_1' \leqslant 0$, and v has a maximum which is less than the maximum of $|v_1|$ so $v < 1/(2q-1)$ when $q > 1$. Upon passing through v_1, v enters a $v' < 0$ region and decreases steadily until it either again intersects the v_1 curve or else approaches some limit as $y \to \infty$. In either case $|v|$ does not exceed $1/(2q-1)$, since, in the first case v would have a minimum at the intersection and in the second $v_1 \to 0$ as $y \to \infty$. The same reasoning may be applied to the remaining points of intersection, if any, of the v and v_1 curves.

In order to obtain an inequality for V itself we rewrite (A2-6) as

$$v' = c - (2b + v)v \tag{A2-11}$$

The solution of this equation which behaves like $y/(2q+2)$ as $y \to 0$ also satisfies the relation

$$v(y) = \int_0^y c(x) \exp\left[-\int_x^y [2b(\xi) + v(\xi)]\, d\xi\right] dx.$$

as may be verified by making use of the relations $c(x) \to 1$ as $x \to 0$ and $2b(\xi) \to (2q+1)/\xi$, $v(\xi) \to \xi/(2q+2)$ as $\xi \to 0$. For then

$$-\int_x^y [2b(\xi) + v(\xi)]\, d\xi \to (2q+1) \log x/y$$
$$v(y) \to \int_0^y (x/y)^{2q+1}\, dx = y/(2q+2)$$

Hence, from (A2-8)

$$1(y_1) = \int_0^{y_1} dy \int_0^y c(x) \exp\left[-\int_x^y [2b(\xi) + v(\xi)]\, d\xi\right] dx$$

and

$$| V(y_1) | \leq \int_0^{y_1} dy \int_0^{y} | c(x) | \exp\left[-\int_x^y [2b(\xi) - | v(\xi) |] \, d\xi \right] dx.$$

From $b \geqslant q$ and $| v | < 1/(2q - 1)$ it follows that $2b(\xi) - | v(\xi) | > 2q - 1$ when $q > 1$. This and $| c(x) | \leqslant (4 + x^2)(1 + x^2)^{-2}/4$ gives

$$| V(y_1) | < \int_0^{\infty} dy \int_0^{y} (4 + x^2)(1 + x^2)^{-2} 4^{-1} \exp[-(2q - 1)(y - x)] \, dx$$

$$= \frac{5\pi}{16(2q - 1)} < \frac{1}{2q - 1}$$

which is the result we set out to establish. The double integral may be reduced to a single integral by inverting the order of integration and integrating with respect to y. Incidentally, most of the roughness of our result is due to the use of the inequality for $| c(x) |$.

References

1. C. E. Shannon, A Mathematical Theory of Communication, *Bell Sys. Tech. Jour.*, 27, 379–423, 623–656 (1948) See especially Section 24.
2. C. E. Shannon, Communication in the Presence of Noise *Proc. I.R.E.*, 37, 10–21 (1949).
3. W. G. Tuller, Theoretical Limitations on the Rate of Transmission of Information *Proc. I.R.E.*, 37, 468–478 (1949).
4. N. Wiener, Cybernetics, John Wiley and Sons (1948).
5. S. Goldman, Some Fundamental Considerations Concerning Noise Reduction and Range in Radar and Communication, *Proc. I.R.E.*, 36, 584–594 (1948).
6. G. N. Watson, Theory of Bessel Functions, Cambridge University Press (1944), equation (1) p. 181.
7. R. A. Fisher, The General Sampling Distribution of the Multiple Correlation Coefficient. *Proc. Roy. Soc. of London* (A) Vol. 121, 654–673 (1928). See in particular pages 669–670.
8. Reference (6), equation (4) p. 394.
9. P. K. Bose, On Recursion Formulae, Tables and Bessel Function Populations Associated with the Distribution of Classical D^2—Statistic, Sankhyā, 8, 235–248 (1947).
10. Compare with §8.4 of reference (6).
11. Reference (6), p. 236.
12. Reference (6), p. 227.

Realization of a Constant Phase Difference

By SIDNEY DARLINGTON

This paper bears on the problem of splitting a signal into two parts of like amplitudes but different phases. Constant phase differences are utilized in such circuits as Hartley single sideband modulators. The networks considered here are pairs of constant-resistance phase-shifting networks connected in parallel at one end. The first part of the paper shows how to compute the best approximation to a constant phase difference obtainable over a prescribed frequency range with a network of prescribed complexity. The latter part shows how to design networks producing the best approximation.

A PERENNIAL problem is that of designing a circuit to split a signal into two parts which are the same in amplitude but which differ in phase by a constant amount. A 90-degree phase difference is needed, for example, in the single sideband modulation system due to R. V. L. Hartley.[1] It is well known that it is not possible to obtain exactly equal amplitudes and exactly constant phase differences at all frequencies except in the trivial special case of a 180-degree phase difference. Various methods have been devised, however, for approximating these characteristics over finite frequency ranges. The most obvious method is to use a pair of constant resistance phase shifting sections in parallel at one end and with separate terminations at the other end[2] as indicated in Fig. 1.

This paper is devoted to the problem of obtaining approximately constant phase differences under the specific assumption that pairs of constant resistance phase shifting networks are to be used. The paper has been written with two objects in mind. The first is the development of a method for determining the best approximation to a constant phase difference which can be obtained over a prescribed frequency range with a pair of phase shifting networks of a prescribed total complexity. The second object is the description of a straightforward design procedure by means of which the networks can be designed to give this best possible approximation.

The problem under consideration is typical of those usually described as problems in network synthesis. In other words, a network of a prescribed general type is to be designed to approximate as closely as possible an ideal operating characteristic of a prescribed form. The same procedure will be followed as that appropriate for most such problems. The procedure begins with the development of a mathematical expression representing the most

[1] *U. S. Patent* 1,666,206, 4/17/28, Modulation System.

[2] Another common method uses reactance shunt branches between effectively infinite impedances, such as the plate and grid impedances of screen grid tubes.

general characteristics which can be obtained with the prescribed type of network. This is followed by the determination of particular choices of the arbitrary constants in the expression, which will lead to the best approximation to the prescribed ideal characteristic. The next step is to determine formulae for the degree of approximation to the ideal, which will be

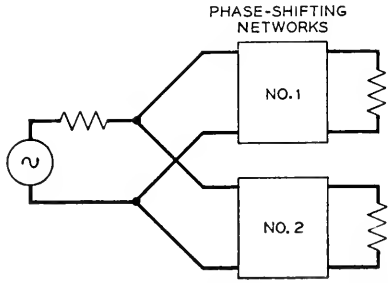

Fig. 1—Phase-shifting networks for approximation to a constant phase difference.

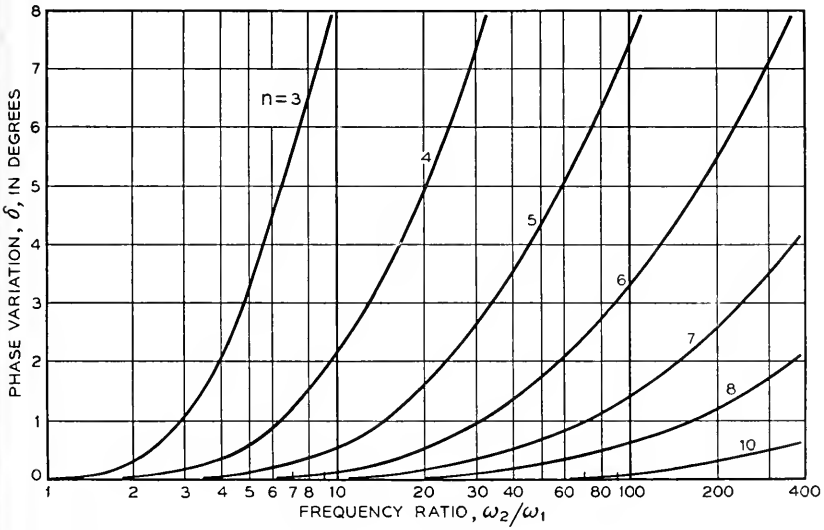

Fig. 2—Variation in phase difference, when average is 90°, with a network of n sections.

obtained with those particular values of the constants. The final step is the development of a method for determining corresponding actual networks.

From the optimum choice of constants, curves can be calculated which show what can be done with a network of any given complexity (Fig. 2). Then the complexity needed for any particular application can be read directly from the curves. The special choice of constants also leads to special

formulae for element values of corresponding networks, using tandem sections of the simplest all-pass type (Fig. 3).

Form of the $\tan\left(\frac{\beta}{2}\right)$ Function

If β_1 and β_2 represent the phase shifts through the two constant resistance networks of Fig. 1, then $\tan\left(\frac{\beta_1}{2}\right)$ and $\tan\left(\frac{\beta_2}{2}\right)$ must both be realizable as the reactances of physical reactance networks. In other words, these quantities must be odd rational functions of ω with real coefficients and must also meet various other special restrictions. If β is used to represent the phase difference $\beta_2 - \beta_1$, the function $\tan\left(\frac{\beta}{2}\right)$ must also be an odd rational function of ω with real coefficients. Because of the minus sign

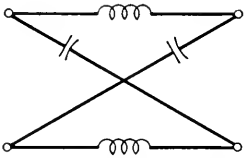

Fig. 3—Simplest all-pass section.

associated with β_1 in the definition of β, however, $\tan\left(\frac{\beta}{2}\right)$ does not have to meet the additional restrictions which must be imposed upon $\tan\left(\frac{\beta_1}{2}\right)$ and $\tan\left(\frac{\beta_2}{2}\right)$. In a later part of the paper a method will be described by which a pair of physical phase shifting networks can be designed to produce any $\tan\left(\frac{\beta}{2}\right)$ function which is an odd rational function of ω with real coefficients.

In any range where the phase difference β approximates a constant, the function $\tan\left(\frac{\beta}{2}\right)$ will also approximate a constant. Hence, the present problem is really that of approximating a constant over a given frequency range with an odd rational function of ω with real coefficients. In this problem, the degree of the function must be assumed to be prescribed as well as the frequency range in which a good approximation is to be obtained, for the degree of the function determines the complexity of the corresponding network.

W. Cauer shows how functions of certain types can be designed to approx-

imate unity in prescribed frequency ranges.[3] These functions, however, are not odd rational functions of frequency but are irrational functions appropriate to represent filter image impedances or the hyperbolic tangents or cotangents of filter transfer constants. It turns out, however, that they can be transformed into odd rational functions of the desired type by a simple transformation of the variable.

Each of Cauer's functions is said to approximate a constant in the Tchebycheff sense, which means that in the prescribed range of good approximation the maximum departure from the approximated constant is as small as is permitted by the specifications on the frequency range and the degree of the function. Each function also has the property of exhibiting series of equal maxima and equal minima in the range of good approximation, such as those indicated in the illustrative β curve[4] of Fig. 4.

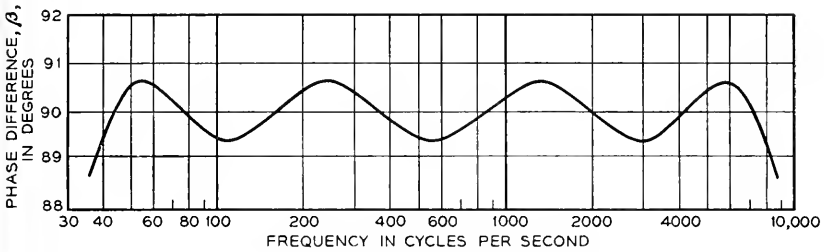

Fig. 4—Example of a phase difference characteristic.

Of the various forms in which Cauer's Tchebycheff functions F can be expressed, the following form is the one appropriate for showing how odd rational functions of frequency can be obtained:

$$(1)\qquad \begin{cases} \text{When } n \text{ is odd} \\ F = U\sqrt{1 - X^2} \prod_{s=1}^{s=\frac{n-1}{2}} \dfrac{\left[1 - sn^2\left(\dfrac{2s-1}{n}K, k\right)X^2\right]}{\left[1 - sn^2\left(\dfrac{2s}{n}K, k\right)X^2\right]}. \\ \text{When } n \text{ is even} \\ F = \dfrac{U}{\sqrt{1 - X^2}} \dfrac{\prod_{s=1}^{s=\frac{n}{2}}\left[1 - sn^2\left(\dfrac{2s-1}{n}K, k\right)X^2\right]}{\prod_{s=1}^{s=\frac{n}{2}-1}\left[1 - sn^2\left(\dfrac{2s}{n}K, k\right)X^2\right]}. \end{cases}$$

[3] "Ein Interpolationsproblem mit Funktionen mit Positivem Realteil," *Mathematische Zeitschrift*, 38, 1–44 (1933).

[4] The data for the illustrative curve were obtained from a trial design carried out by P. W. Rounds.

In these equations, the symbol *sn* indicates an elliptic sine, of modulus k, while K represents the corresponding complete elliptic integral. U is merely a constant scale factor, while n is an integer measuring the complexity of corresponding networks. In the case of phase-difference networks, n represents the total number of sections of the type indicated in Fig. 3, which are included in the two phase-shifting networks or their tandem section equivalents.

In Cauer's filter theory, the variable X represents a rational function of ω which permits F to be an image impedance or a $\coth\left(\frac{\theta}{2}\right)$ function. In order that F may be an odd rational function of ω, however, as is required when it is to represent $\tan\left(\frac{\beta}{2}\right)$, X must be defined by the relation

$$\omega = \omega_2\sqrt{1 - X^2}. \tag{2}$$

Cauer shows that F approximates a constant in the Tchebycheff sense in the range $0 < X < k$. Hence, in terms of ω, the range of approximation is $\omega_1 < \omega < \omega_2$, where ω_1 and ω_2 are arbitrary provided the modulus k is assumed to be determined by the relation

$$k = \frac{\sqrt{\omega_2^2 - \omega_1^2}}{\omega_2}. \tag{3}$$

Alternative Expression for the $\tan\left(\frac{\beta}{2}\right)$ Function

While equations (1) are the most convenient form of F to use in deriving the transformation of the variable, an alternative more compact form is more suitable for determining the degree of approximation to a constant phase difference and the element values of corresponding networks. When F represents $\tan\left(\frac{\beta}{2}\right)$ and hence ω and X are related as in (2), the equivalent expression is as follows:[5]

$$\tan\left(\frac{\beta}{2}\right) = U\, dn\left(nu\frac{K_1}{K}, k_1\right)$$
$$\omega = \omega_2\, dn(u, k). \tag{4}$$

In this expression, *dn* represents a so-called "*dn*" function, the third type of Jacobian elliptic function usually associated with the elliptic sine, or *sn* function, and the elliptic cosine, or *cn* function. The symbol u represents

[5] This expression depends on a so called modular transformation of elliptic functions not found in the usual elliptic function text. The transformation theory may be found in "An Elementary Treatise on Elliptic Functions," Arthur Cayley, G. Bell & Sons, London, 1895.

a "parametric variable" which would be eliminated on forming a single equation from the two simultaneous equations indicated. The modulus k_1, of the *dn* function corresponding to $\tan\left(\frac{\beta}{2}\right)$ is related to the modulus k, of the *dn* function corresponding to ω, in the manner indicated below. The constant K_1, of course, represents the complete integral of modulus k_1, just as K represents the complete integral of modulus k.

Corresponding to any modulus k there is a so-called modular constant q. Using q_1 to represent the corresponding modular constant of modulus k_1, it is here required that

$$q_1 = q^n. \tag{5}$$

One modulus can be computed from the other by means of this relationship and tabulations of $\log_{10} q$ vs $\sin^{-1} k$ which are included in most elliptic function tables.[6]

Degree of Approximation to a Constant Phase Difference

When u is real and varies from zero to infinity, the corresponding value of ω as determined by (4) merely oscillates back and forth between the values ω_1 and ω_2. In other words, it merely crosses back and forth across the range in which $\tan\left(\frac{\beta}{2}\right)$ approximates a constant. Similarly, when u is real and increases from zero to infinity, $\tan\left(\frac{\beta}{2}\right)$ oscillates between $U\sqrt{1-k_1^2}$ and U. The *equal ripple* property of the curve illustrated in Fig. 4 is explained by the fact that the period of oscillation of $\tan\left(\frac{\beta}{2}\right)$ with respect to u is merely a fraction of that of ω, so that $\tan\left(\frac{\beta}{2}\right)$ passes through several ripples while the value of ω moves from ω_1 to ω_2.

Combining the formulae for the maximum and minimum values of $\tan\left(\frac{\beta}{2}\right)$ gives the relation

$$\tan\left(\frac{\delta}{2}\right) = \frac{U\left(1-\sqrt{1-k_1^2}\right)}{1+U^2\sqrt{1-k_1^2}} \tag{6}$$

[6] When k is extremely close to unity, it may be easier to obtain accurate computations by using the additional relation

$$\log_{10}(q)\,\log_{10}(q') = \left(\frac{\pi}{\log_e(10)}\right)^2$$

where q' is the modular constant of modulus $\sqrt{1-k^2} = \frac{\omega_1}{\omega_2}$.

in which δ represents the total variation of the phase difference β in the approximation range. Similarly, the average value β_a of β in the approximation range is given by[7]

$$\tan(\beta_a) = \frac{U(1 + \sqrt{1 - k_1^2})}{1 - U^2\sqrt{1 - k_1^2}}. \tag{7}$$

If the phase variation δ is reasonably small, (6) and (7) can be replaced by the approximate relationships

$$\begin{aligned} \delta &= \frac{\sin(\beta_a)}{2} k_1^2 \quad \text{radians} \\ \tan\left(\frac{\beta_a}{2}\right) &= U\sqrt[4]{1 - k_1^2}.^{8} \end{aligned} \tag{8}$$

A still further modification is obtained by replacing k_1^2 by the quantity $16q_1$, which is an approximate equivalent when k_1^2 is small, and by then replacing q_1 by the equivalent q^n of (5). This gives

$$\begin{aligned} \delta &= 8\sin(\beta_a)q^n \\ \tan\left(\frac{\beta_a}{2}\right) &= U\sqrt[4]{1 - 16q^n}. \end{aligned} \tag{9}$$

When combined with (3) and tabulations of $\sin^{-1}(k)$ vs $\log_{10}(q)$, these formulae can be used to compute δ when the parameters ω_1, ω_2, β_a and n are prescribed. Curves of δ are plotted against ω_2 ω_1 in Fig. 2, assuming β_a to be 90 degrees.

Determination of a Network Corresponding to a General Phase Difference Function

Since $\tan\left(\frac{\beta}{2}\right)$ must be an odd rational function of ω, it can be expressed in the form

$$\tan\left(\frac{\beta}{2}\right) = \frac{\omega B}{A} \tag{10}$$

in which A and B are even polynomials in ω. This requires

$$\frac{\beta}{2} = \arg(A + i\omega B). \tag{11}$$

[7] More exactly, β_a is the average of the maximum and minimum values of β occurring in the range of approximation.

[8] In the important special case in which the average phase difference β_a is 90°, this expression for $\tan\left(\frac{\beta_a}{2}\right)$ is exact rather than approximate.

Similarly, if attention is focused on the phase shifts of the individual phase-shifting networks rather than on the phase difference, the following odd rational functions can be introduced:

$$\tan\left(\frac{\beta_1}{2}\right) = \frac{\omega B_1}{A_1}$$
$$\tan\left(\frac{\beta_2}{2}\right) = \frac{\omega B_2}{A_2} \tag{12}$$

in which A_1, B_1, A_2, and B_2 are additional even polynomials in ω. This requires

$$\frac{\beta_1}{2} = \arg\,(A_1 + i\omega B_1)$$
$$\frac{\beta_2}{2} = \arg\,(A_2 + i\omega\beta_2). \tag{13}$$

It also requires

$$-\frac{\beta_1}{2} = \arg\,(A_1 - i\omega B_1). \tag{14}$$

Since the argument of a product is the sum of the arguments of the separate factors, (13) and (14) require

$$\frac{\beta}{2} = \frac{\beta_2 - \beta_1}{2} = \arg\,(A_2 + i\omega B_2)(A_1 - i\omega B_1). \tag{15}$$

This permits us to write

$$(A_2 + i\omega B_2)(A_1 - i\omega B_1) = H(A + i\omega B) \tag{16}$$

in which H is a real constant.

When $\tan\left(\frac{\beta}{2}\right)$ is prescribed, a corresponding polynomial of the form $(A + i\omega B)$ can readily be derived. The problem is then to factor it into the product of two polynomials $(A_2 + i\omega B_2)$ and $(A_1 - i\omega B_1)$ such that A_1, B_1, A_2, and B_2 determine physically realizable phase shifts through (12). Two factors of the general form $(A_2 + i\omega B_2)$ and $(A_1 - i\omega B_1)$ can readily be obtained in a number of ways. The only question is how to obtain them in such a way that the corresponding phase characteristics will be physical. A procedure meeting this requirement is described below.

The variable ω is first replaced in $(A + i\omega B)$ by p representing $i\omega$. This leaves a polynomial in p with real coefficients, since A and B represent polynomials in ω^2, while p^2 represents $-\omega^2$. Suppose all the roots of the polynomial $A + pB$ are determined. Then this polynomial can be split into

two factors by assigning various of the roots to each of the two factors. It turns out that physically realizable phase characteristics will be obtained if all those roots with positive real parts are assigned to the factor $(A_1 - pB_1)$ which appears in (16) when $i\omega$ is replaced by p, all other roots being assigned to the factor $(A_2 + pB_2)$.

The physical realizability of the above division of the roots follows from a theorem which states that $\frac{pB_x}{A_x}$ is realizable as the impedance of a two-terminal reactance network whenever A_x and B_x are even polynomials in p with real coefficients such that $A_x + pB_x$ has no roots with positive real parts.[9] From this theorem and the fact that the evenness of A_x and B_x causes them to remain unchanged when p is reversed in sign, it follows that $\frac{pB_x}{A_x}$ will also be the impedance of a physical two-terminal reactance network whenever $A_x - pB_x$ has no roots with *negative* real parts. Thus, by (12) the above division of the roots of $A + pB$ makes $\tan\left(\frac{\beta_1}{2}\right)$ and $\tan\left(\frac{\beta_2}{2}\right)$ realizable as the impedances of two-terminal reactance networks. These reactance networks and their inverses are merely the arms of unit impedance lattices producing the phase characteristics defined by (12).

The above argument merely shows that each of the two phase-shifting networks can at least be realized as a single lattice when $\tan\left(\frac{\beta_1}{2}\right)$ and $\tan\left(\frac{\beta_2}{2}\right)$ are determined by the method described. Actually, they can be broken into tandem sections directly as soon as the roots of $(A_1 - pB_1)$ and $(A_2 + pB_2)$ have been determined. From $(A_1 - pB_1)$, the quantity $(A_1 + pB_1)$ can be found by merely reversing the signs of the roots. Then by using the principle that the argument of a product is the sum of the arguments of the separate factors, phase-shifting networks can be designed corresponding to various factors or groups of factors as determined from the known roots of $(A_1 + pB_1)$ and $(A_2 + pB_2)$. There can be a separate section for each real root and each conjugate pair of complex roots.[10]

Determination of a Network Corresponding to a Tchebycheff Type of Phase Difference Characteristic

The procedure described above for determining a network corresponding to a general phase difference characteristic is complicated by the necessity

[9] See "Synthesis of Reactance 4-Poles which Produce Prescribed Insertion Loss Characteristics," *Journal of Mathematics and Physics*, Vol. XVIII, No. 4, September, 1939—page 276.

[10] See H. W. Bode, "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Company, New York, 1945, Page 239, §11.6.

of determining the roots of the polynomial $A + pB$. In the case of the Tchebycheff type of characteristic described in the first part of the paper, the required roots can be determined by means of special relationships. In the first place, the roots of $A + pB$ are the roots of $\left(1 + i \tan \frac{\beta}{2}\right)$. In other words, by equation (4) they are the roots of $\left[1 + iU\, dn\left(nu \frac{K_1}{K}, k_1\right)\right]$. The values of u at the roots turn out to have an imaginary part iK', where K' is the complete elliptic integral of modulus $\sqrt{1 - k^2}$. If a new variable u' is defined by

$$u = u' + iK' \tag{17}$$

the roots can be shown to correspond to the values of u' determined by

$$\frac{sn\left(nu' \frac{K_1}{K}, k_1\right)}{cn\left(nu' \frac{K_1}{K}, k_1\right)} = -U. \tag{18}$$

If it is assumed that the phase variation is small in the range of approximation to a constant, it can be shown that one value of u' determined by the above relation is given approximately by

$$\frac{nu'\pi}{K} = -\beta_a \tag{19}$$

where β_a is the average phase difference for the range of approximation as before (in radians). After this value of u' has been computed, all the roots of $\left[1 + iU\, dn\left(nu \frac{K_1}{K}, k_1\right)\right]$ can be found by computing the values of ω corresponding to this value of u' and to those values obtained by adding integral multiples of the real period $\frac{2K}{n}$ of $dn\left(nu \frac{K_1}{K}, k_1\right)$. This gives the following formula for the roots in terms of $p = i\omega$.

$$p_\sigma = \omega_2 \frac{cn\left(\frac{2\sigma K}{n} + u_0'\right)}{sn\left(\frac{2\sigma K}{n} + u_0'\right)}, \qquad \sigma = 0, \cdots, (n-1) \tag{20}$$

in which u_0' is the value of u' determined by (19).

Finally, instead of using the above elliptic function formula directly, one may replace the elliptic functions by equivalent ratios of Fourier series expansions of θ functions. This gives

$$p_\sigma = \sqrt{\omega_1\omega_2}\, \frac{\cos(\lambda_\sigma) + q^2 \cos(3\lambda_\sigma) + q^6 \cos(5\lambda_\sigma) \cdots}{\sin(\lambda_\sigma) - q^2 \sin(3\lambda_\sigma) + q^6 \sin(5\lambda_\sigma) \cdots} \tag{21}$$

in which the angle λ_σ is defined by

$$\lambda_\sigma = \frac{\sigma \cdot 180^\circ - \frac{1}{2}\beta_a}{n} \text{ degrees}, \qquad \sigma = 0, \cdots, (n-1). \tag{22}$$

Because all the p_σ's are real in this Tchebycheff case, corresponding networks can be made up of sections of the simple type indicated in Fig. 3. In one of the two phase-shifting networks there will be one section for each positive p_σ, and it will be given by

$$L = \frac{R_0}{p_\sigma} \qquad C = \frac{1}{R_0 p_\sigma}$$

where R_0 is the image impedance. Similarly, in the second phase-shifting network there will be one section for each negative p_σ, and it will be given by

$$L = -\frac{R_0}{p_\sigma} \qquad C = \frac{-1}{R_0 p_\sigma}.$$

Conversion of Concentrated Loads on Wood Crossarms to Loads Distributed at Each Pin Position

By RICHARD C. EGGLESTON

ONE of the most important requisites in all fields of engineering endeavor is knowledge of the strength of materials. The development of testing machines and techniques to study the basic properties of metals, plastics and wood products to withstand breaking forces has been a distinctive achievement during the last half century. All materials, whether they be part of a bridge, a building, a shipping crate, a telephone pole or a crossarm on a telephone pole, break under an excessive stress. To have accurate knowledge of the strength of the millions of crossarms used to carry the regular load of wires, which are frequently subjected to the extra loads of wind and ice, is most important in electrical communication.

When strength tests of crossarms are made, the information most generally sought is how great a vertical load equally distributed at each insulator pin hole will the arms stand. In the past many crossarm tests have been made by the concentrated load method, where the arm is either supported at each end and loaded at the center, or supported at the center and loaded at the ends until failure occurs (Fig. 1, a and b). Some have been made by the distributed load method by placing, manually and simultaneously, 50-pound weights in wire baskets suspended from each pin hole, and continuing such load applications until the arm fails. The method is objectionable chiefly because, in many of the tests, the loading is inadvertently carried past the maximum loads the arms will support. This objection was overcome in recent tests made by the Bell Telephone Laboratories[1], where the loads were also distributed at each pin position. However, instead of subjecting the 10-pin test arms to sudden 500-pound load increments (viz. 50 pounds at each of the 10 pin holes), the loads were applied gradually by a hydraulic testing machine (Fig. 1, c). But, in spite of the advantages of this machine method of distributed load application, it is probable that, because of the less elaborate apparatus involved in simple beam tests, there will continue to be tests made by the concentrated load method.

Where tests have been made by the concentrated load method, the question arises how can the results be converted to a load-per-pin basis? A conversion is needed before a fair comparison can be made of all test results, and also to furnish the information generally most wanted, which is, as

[1]Bell System Monograph No. B-1563, Strength Tests of Wood Crossarms.

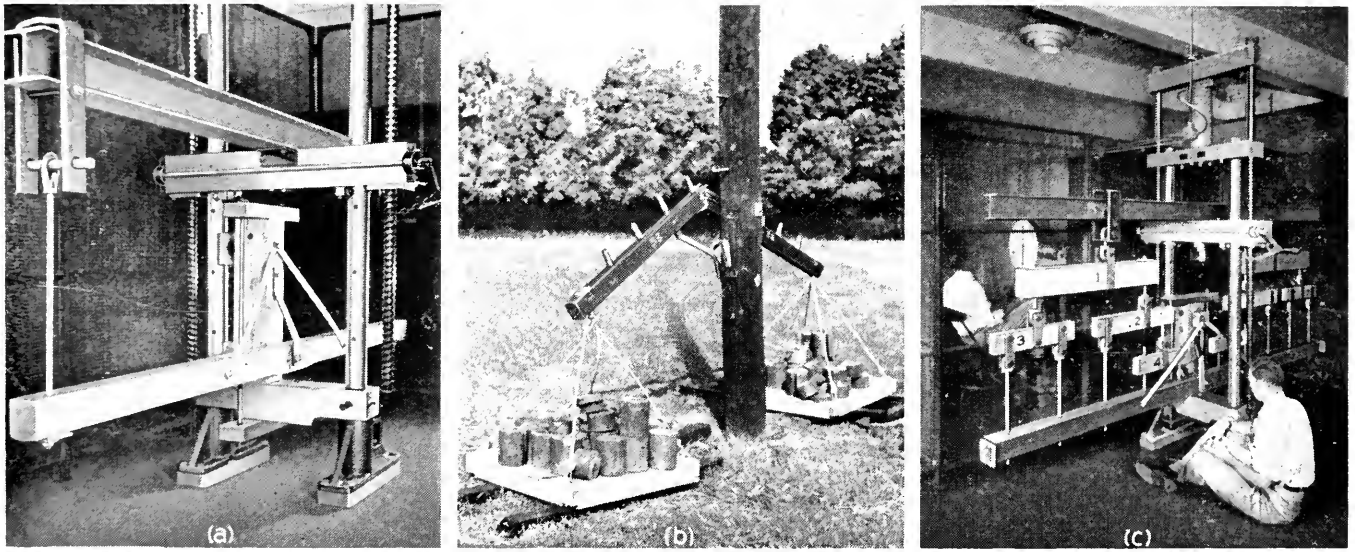

Fig. 1—Photographs of crossarm tests:
a—Arm supported at center in testing machine ready for application of load at end pin holes.
b—Arm supported at center on a pole. This roofed 10A arm was loaded manually at end pin holes and failed at a pole pin hole (critical) section.
c—Arm supported at center in testing machine ready for application of load at each pin position.

previously stated, what load per pin will the arm support? There are more than twenty million crossarms in the pole lines of the Bell System and each year about a million arms are added. A complete understanding of every problem associated with this important item of outside plant material is manifestly worth while. This paper is intended to contribute to that end. It presents a solution of the problem of converting concentrated vertical loads to comparable loads distributed at each insulator pin position.

The location of the critical section in crossarms is a basic factor in a study of the problem. The critical section of a crossarm is the section at which the fiber stress is greatest when the arm is loaded. It is the section where the arm may be expected to break if overloaded. To determine its location, the bending moment at various sections along the arm is divided by the section modulus of the respective sections. The quotient in each instance is the fiber stress for each section investigated. The location showing the greatest fiber stress is the critical section. Since horizontal shear is not the controlling stress in crossarm failures under loads distributed at each pin hole, bending stresses only were considered in this analysis.

Because of the differences in arm shape and in the spacing of pin holes, the location of the critical section is not the same in all arms. It is estimated that at least three fourths of the arms in the Bell System are 10A and 10B crossarms.[2] Both are 10 feet long and 3.25″ x 4.25″ in cross section. In the 10A arm (Fig. 3), the space between the pin holes is 12 inches, except between the pole pin holes, where the space is 16 inches. In the 10B (Fig. 4) the pin hole spacing is 10 inches with a 32-inch space between the pole pins. Both types are bored for wood pins. Most of the arms now in the plant are "roofed", that is, the top surface of the arm, except the center foot of length, is rounded on a radius of about 4.25 inches. Under the current design, however, the top surface of Bell System arms is flat, except for the edges, which are beveled.

Previous studies of both roofed and beveled arms of various types have shown that the critical section of clear arms under vertical loads is either at the center or at the pole pin hole sections. This study is confined to those sections of clear 10A and 10B crossarms of nominal dimensions, both roofed and beveled. Moreover, it was assumed for the purpose of load analysis, that the crossarms are supported at the center only; since, under loads on each side of the pole, the standard crossarm braces provide no significant support when the loads are sufficient to break the arm.

Roofed 10A Arm

Let it be assumed that the breaking load concentrated in each end pin hole of a roofed 10A arm is 800 pounds. As shown in Calculation 1 in the

[2]10A and 10B crossarms were formerly known as Type A and Type B crossarms, respectively.

appendix, the bending moment at the center of the arm from the assumed loads would be 44,800 pound-inches, and the fiber stress at the center would be 4660 psi. That calculation also shows that the bending moment at the pole pin holes would be 38,400 pound-inches, and the fiber stress at the pole pin holes 7515 psi. Since the stress at the pole pin holes is greater than that at the center, the critical section of a roofed 10A arm is at the pole pin holes when the arm is subjected to a breaking load at each end pin hole.

The information wanted, however, is what load at each of the ten pin positions would have produced the same moment and same fiber stress at the critical section? A *tentative* answer is found by dividing the 38,400 bending moment by the "total-per-pin" lever arm[3], 120″ (see Calculation 1) or 320 pounds. Checking to determine whether the location of the critical section changes under loads of 320 pounds at each pin position, Calculation 1 shows that the fiber stress at the pole pin holes and at the center would be 7515 psi and 5257 psi, respectively. Since the stress at the pole pin holes is greater, it is clear that the critical section is there also under equal loads at each pin position; and the 320-pound load per pin is comparable to the concentrated load of 800 pounds at each end of the arm.

If a similar investigation were made of a roofed 10B arm and of a beveled 10A arm, it would be found that the pole pin hole section is the critical section of these arms; and that the load per pin comparable to concentrated loads at the arm ends would, like the roofed 10A arm, be equal to the bending moment at the pole pin hole section due to the concentrated load divided by the total per pin lever arm to that section. Figure 1b shows a roofed 10A arm that broke under test at a pole pin hole (critical) section from concentrated loads at the ends of the arm.

Beveled 10B Arm

For the investigation of this arm, let a breaking load of 1000 pounds at each end pin hole be assumed. Incidentally, it should be noted that so far as this analysis is concerned, the magnitude of the assumed concentrated loads is of no importance. However, since both computations and tests show the 10B arm to be stronger than the 10A, it seemed appropriate to assume a larger concentrated load for the 10B arm.

As shown in Calculation 2 of the appendix, the bending moment at the center due to the 1000-pound load would be 56,000 pound-inches and the fiber stress 5882 psi, while at the pole pin holes the bending moment would be 40,000 pound-inches and the fiber stress 6885 psi. Here again, under concentrated loads at each end pin hole, the critical section is at the pole pin holes.

[3]By total-per-pin lever arm is meant the summation of the distances from each pin position to the section concerned—in this instance to the pole pin hole section.

Calculating the load for each of the 10 pin positions in the same manner as for the roofed 10A arm, we have, tentatively, a load per pin of 400 pounds. However, in checking to determine whether the location of the critical section changes under loads of 400 pounds at each pin position, we obtain results quite different from those in Calculation 1; for Calculation 2 indicates a fiber stress of 6885 psi at the pole pin holes, but a higher stress (7563 psi) at the center, which shows that the location of the critical section does change. Moreover, this change would occur whether the loads were 400 pounds per pin or 4 pounds per pin. But let us now consider the 400-pound load.

If a concentrated load of 1000 pounds results in a fiber stress at the pole pin hole section of 6885 psi and causes failure, that stress is the *maximum* ultimate fiber stress for the arm. It is, therefore, not reasonable to suppose that the same arm would have endured a higher stress (7563 psi) at the center if it had been loaded at each pin position. If 6885 psi is the maximum stress for the arm, the maximum moment it would endure at its center would be 65,500 pound-inches (viz. 6885 multiplied by 9.52, the section modulus of the center section). The maximum load per pin would be 364 pounds (viz. 65,500 divided by 180, the total-per-pin lever arm to the center); and this load of 364 pounds, not 400 pounds, distributed at the 10 pin positions is comparable to the 1000-pound concentrated load. Thus, while the critical section of a beveled 10B arm is at the pole pin holes when the load is concentrated at the arm ends, it shifts to the center when the load is distributed at each pin position; and, moreover, the load is less than the load per pin tentatively computed.

A graphic illustration of this shift of the critical section is shown in Fig. 2. Graph 1 in this figure is the graph of the resisting moments of a clear, straight-grained beveled 10B arm, 3.25″ x 4.25″ in cross-section, and having an assumed ultimate fiber strength in bending of 6000 psi. Each point in the graph is equal to the section modulus of the section under consideration multiplied by 6000 psi. Graph 2, which is the graph of a concentrated load at the end pin position, was drawn from the zero moment under the end pin to the point of greatest moment possible without intersecting resisting moment Graph 1. Since the point of coincidence between Graphs 1 and 2 is the pole pin hole section, that section is the critical section for a concentrated load at the end pin. The magnitude of this concentrated load is equal to the resisting moment at the pole pin hole, 34,860 pound-inches (viz. 5.81 inches^3 x 6000 psi) divided by the 40″ lever arm, or 871.5 pounds. The load per pin, *tentatively figured*, would be 34,860 pound-inches divided by 100 inches or 348.6 pounds. Graph 3 is the graph of a load (P) of 348.6 pounds at each pin hole. Under such loading, however, the bending moment at the center of the crossarm would be 62,748 pound-

inches (viz. 348.6 x 180), which exceeds the 57,120 pound-inches resisting moment at the center (viz. 9.52 inches^3 x 6000 psi). This means that the

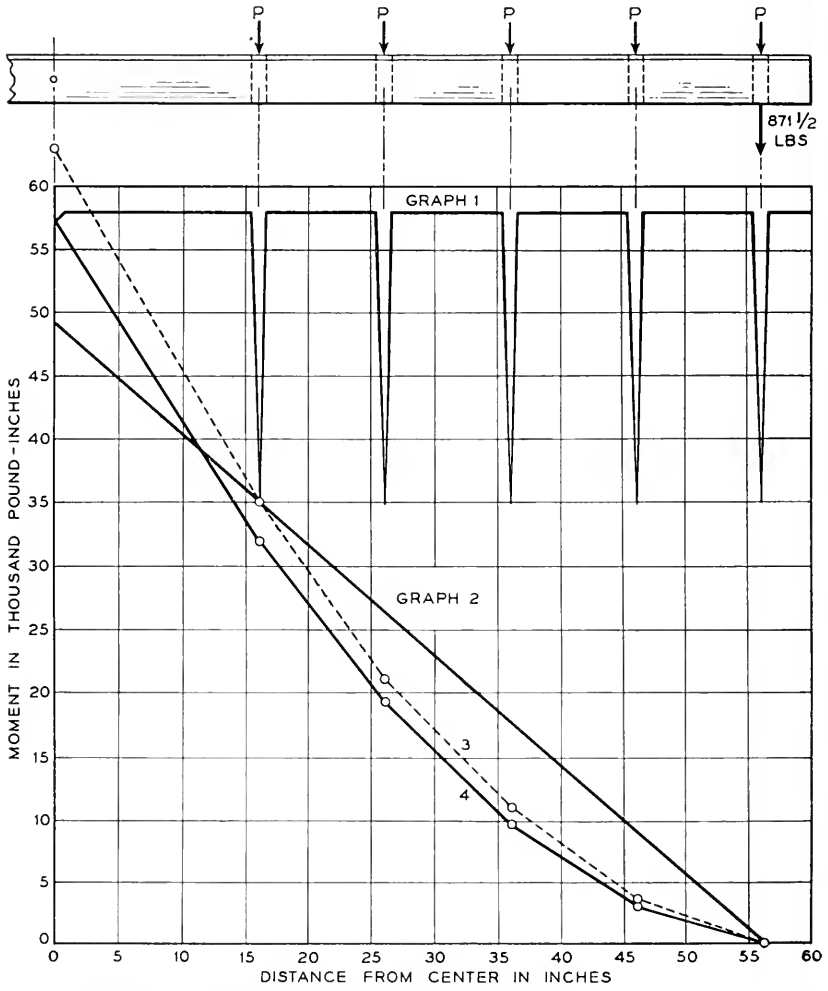

Fig. 2—Moment diagrams for a beveled 10B crossarm:
Graph 1—Resisting moments of a clear, straight grained, 3.25″ x 4.25″ arm. Fiber stress assumed to be 6000 psi;
Graph 2—Bending moments from a concentrated load of 871.5 pounds at end pin hole;
Graph 3—Bending moments from a load of 348.6 pounds at each pin hole; and
Graph 4—Bending moments from a load of 317.3 pounds at each pin hole.

arm would fail under such loading; and that the critical section of the arm under loads distributed at each pin hole is not at the pole pin holes but at the center of the arm. The maximum load per pin that the arm would

endure is the resisting moment at the center divided by the total-per-pin lever arm, or 57,120 pound-inches divided by 180 inches or 317.3 pounds. Graph 4 is the bending moment graph of this 317.3-pound maximum load per pin.

SUMMARY

Let W = Concentrated load,
P = Load per pin,
Mp = Bending moment at pole pin hole section,
fc = Fiber stress at center section,
fp = Fiber stress at pole pin hole section,
Sc = Section modulus of center section, and
Sp = Section modulus of pole pin hole section.

Using this notation, the results of the analyses may be summarized as follows:

For 10A arms both roofed and beveled:

$$Mp = 48W \quad \text{(for concentrated loads), and}$$

$$Mp = 120P \quad \text{(for pin loads). Therefore,}$$

$$P = \frac{48W}{120} = 0.4W$$

For 10B arms-roofed:

$$Mp = 40W \quad \text{(for concentrated loads), and}$$

$$Mp = 100P \quad \text{(for pin loads). Therefore,}$$

$$P = \frac{40W}{100} = 0.4W$$

For the beveled 10B arm, however, where the critical section is at the center, the value P = 0.4W does not apply. The value of P would be such as to produce the same fiber stress at the center section as the fiber stress resulting from the concentrated load (W) at the pole pin hole section. Thus

$$fc = \frac{180P}{Sc} \quad \text{and}$$

$$fp = \frac{40W}{Sp}$$

Equating these, we have

$$\frac{180P}{Sc} = \frac{40W}{Sp} \quad \text{and}$$

$$P = \frac{40W}{Sp} \times \frac{Sc}{180} = \frac{2ScW}{9Sp} = \frac{2 \times 9.52W}{9 \times 5.81} = 0.364W$$

Therefore, under the conditions assumed, and only under such conditions, we may say that the loads per pin (P) comparable to the assumed concentrated loads (W) would be

$$P = 0.4W \text{ for } \begin{cases} \text{10A arms—roofed} \\ \text{10A arms—beveled} \\ \text{10B arms—roofed} \end{cases}$$

and

$$P = 0.364W \text{ for 10B arms—beveled}$$

While these results are restricted to the four arm types listed, the same principles followed in arriving at these results may be applied to other types and sizes of arms, and to other conditions of loading. Whether the conversion of single concentrated loads to loads per pin is performed by the method illustrated in Calculations 1 and 2 of the appendix, or is done by a moment diagram, as in Fig. 2, the procedure recommended is as follows:

Step 1. Determine the critical section under the concentrated load.

Step 2. Divide the bending moment at the critical section by the total-per-pin lever arm to the critical section to determine the load per pin.

Step 3. Check the fiber stress (under such loads per pin) at various sections to see whether the *location* of the critical section differs under load per pin.

Step 4. If it does differ, proceed as shown for the beveled 10B arm (viz., the comparable load per pin is equal to the resisting moment of the *critical section* divided by the total-per-pin lever arm to the critical section). If it does not differ, the load per pin as determined in Step 2 is the comparable load per pin sought.

Conclusions

(1) The location of the critical section under loads distributed at each pin position must be determined before undertaking the conversion of concentrated loads to distributed loads.

(2) The location of the critical section of a crossarm under a given condition of loading may or may not be the same under a different condition of loading.

(3) The load per pin comparable with a given concentrated load is equal to the resisting moment of the *critical* section divided by the total-per-pin lever arm to the critical section.

(4) While the results shown are confined to the conversion of concentrated vertical loads to distributed loads for 10A and 10B arms only, the principles of the study may be applied to other types and sizes of arms and to other conditions of loading.

APPENDIX

Calculation 1. Bending Moments and Fiber Stresses in a Roofed 10A Crossarm—(See Figure 3)

Notation:

W = 800 pounds concentrated load
P = Load per pin
Mc = Bending moment at arm center
Mp = Bending moment at pole pin hole
fc = Fiber stress at center
fp = Fiber stress at pole pin hole
Sc = Section modulus of center section[4]
Sp = Section modulus of pole pin hole section[4]

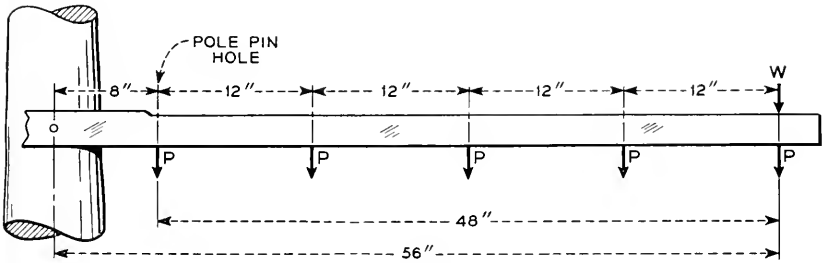

Fig. 3—Loading diagram for a roofed 10A crossarm.

Concentrated Load:

$Mc = W \times 56 = 800 \times 56 = 44{,}800$ pound-inches
$Mp = W \times 48 = 800 \times 48 = 38{,}400$ pound-inches
$fc = Mc \div Sc = 44{,}800 \div 9.74 = 4600$ psi
$fp = Mp \div Sp = 38{,}400 \div 5.11 = 7515$ psi

Load per Pin:

$Mc = 56P + 44P + 32P + 20P + 8P = 160P$
$Mp = 48P + 36P + 24P + 12P = 120P$

(Note: The total-per-pin lever arms are 160″ to center and 120″ to the pole pin hole).

Since under W load f is maximum at pole pin hole, the P load that would result in same f is $P = 38{,}400 \div 120 = 320$ pounds. Thus

$fp = 120P \div Sp = (120 \times 320) \div 5.11 = 7515$ psi
$fc = 160P \div Sc = (160 \times 320) \div 9.74 = 5257$ psi

Conclusion:

Under both W loads and P loads, the critical section is the pole pin hole section.

[4] $Sc = 9.74$ and $Sp = 5.11$ for clear roofed 3.25″ x 4.25″ crossarms. (See Pages 27 and 28 of *Bell Sys. Tech. Jour.*, Jan. 1945).

Calculation 2. Bending Moments and Fiber Stresses in a Beveled 10B Crossarm—(See Figure 4)

Notation:

W = 1000 pounds concentrated load
P = Load per pin
Mc = Bending moment at arm center
Mp = Bending moment at pole pin hole
fc = Fiber stress at center
fp = Fiber stress at pole pin hole
Sc = Section modulus of center section[5]
Sp = Section modulus of pole pin hole section[5]

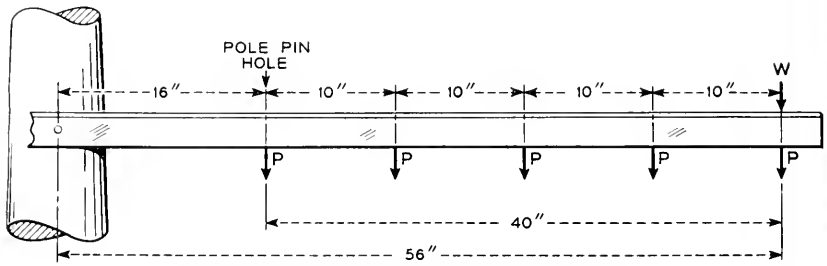

Fig. 4—Loading diagram for a beveled 10B crossarm.

Concentrated Load:

$Mc = W \times 56 = 1000 \times 56 = 56{,}000$ pound-inches
$Mp = W \times 40 = 1000 \times 40 = 40{,}000$ pound-inches
$fc = Mc \div Sc = 56{,}000 \div 9.52 = 5882$ psi
$fp = Mp \div Sp = 40{,}000 \div 5.81 = 6885$ psi

Load per Pin:

$Mc = 56P + 46P + 36P + 26P + 16P = 180P$
$Mp = 40P + 30P + 20P + 10P = 100P$
$P = 40{,}000 \div 100 = 400$ pounds

$$fp = \frac{100P}{Sp} = \frac{100 \times 400}{5.81} = 6885 \text{ psi}$$

$$fc = \frac{180}{Sc} = \frac{180 \times 400}{9.52} = 7563 \text{ psi}$$

Conclusion:

Critical section shifts under P loads, and arm will not support 400 pounds per pin.

[5] $Sc = 9.52$ and $Sp = 5.81$ for clear beveled 3.25" x 4.25" crossarms. (See Calculation 3 of this appendix.)

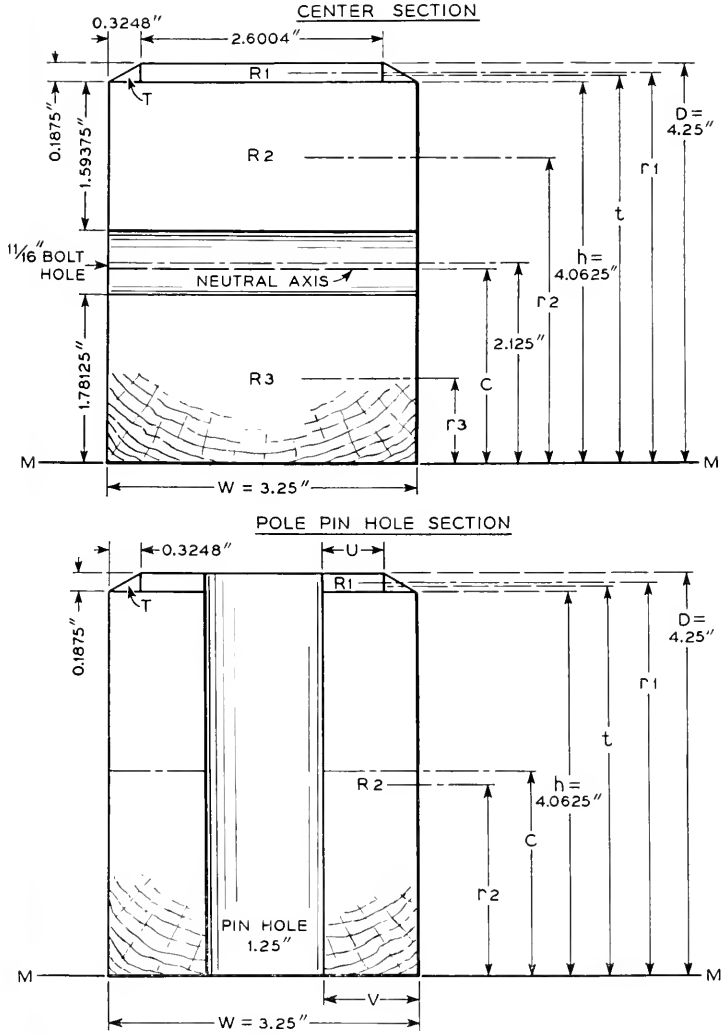

Fig. 5—Beveled crossarm sections showing significance of the notation used in Calculation 3 of this appendix.

Calculation 3. Section Modulus—Clear Beveled Sections
(The notation used in this calculation is shown in Fig. 5)

		Center Section	*Pole Pin Hole Section*
Section width (W)	Inches	3.25	3.25
Section depth (D)	Inches	4.25	4.25
$V = (W - 1.25'') \div 2$	Inches		1.00
$U = V - .3248''$	Inches		.6752
Areas:			
T	Sq. Ins.	$(2T)$ .0609	.0304
$R1$	Sq. Ins.	.4876	.1266
$R2$	Sq. Ins.	5.1797	4.0625
$R3$	Sq. Ins.	5.7891	—
Total 1	Sq. Ins.	11.5173	4.2145
$t = h + (.1875'' \div 3)$	Inches	4.1250	4.1250
$r1 = h + (.1875'' \div 2)$	Inches	4.1563	4.1563
$r2 = h - (1.59375'' \div 2)$	Inches	3.2656	$(\frac{1}{2}h)$ 2.0313
$r3 = 1.78125'' \div 2$	Inches	.8906	—
Moments about MM:			
Tt	Ins.^{3}	$(2Tt)$ .2512	.1254
$R1r1$	Ins.^{3}	2.0266	.5262
$R2r2$	Ins.^{3}	16.9148	8.2522
$R3r3$	Ins.^{3}	5.1558	—
Total 2	Ins.^{3}	24.3484	8.9038
c = Total 2 ÷ Total 1	Inches	2.1141	2.1127
$dt = t - c$	Inches	2.0109	2.0123
$dr1 = r1 - c$	Inches	2.0422	2.0436
$dr2 = r2 - c$	Inches	1.1515	$(c - r2)$ .0812
$dr3 = c - r3$	Inches	1.2151	—
Moments of Inertia:			
IT	Ins.^{4}	$(2IT)$ .0001	.00006
$IR1$	Ins.^{4}	.0014	.0003
$IR2$	Ins.^{4}	1.0964	5.5873
$IR3$	Ins.^{4}	1.5307	—
$T(dt)^2$	Ins.^{4}	$[2T(dt)^2]$ .2463	.0612
$R1(dr1)^2$	Ins.^{4}	2.0336	.5287
$R2(dr2)^2$	Ins.^{4}	6.8680	.0268
$R3(dr3)^2$	Ins.^{4}	8.5474	—
I	Ins.^{4}	20.3239	6.2044
$y = D - c$	Inches	2.1359	2.1373
Section Modulus:			
$S = I \div y$	Ins.^{3}	9.52	$(\frac{1}{2}S)$ 2.9029
			S = 5.81

The Linear Theory of Fluctuations Arising from Diffusional Mechanisms—An Attempt at a Theory of Contact Noise

By J. M. RICHARDSON

The spectral density is calculated for the electrical resistance when it is linearly coupled to a diffusing medium (particles or heat) undergoing thermally excited fluctuations. Specific forms of the spectral density are given for several types of coupling which are simple and physically reasonable. The principal objective is the understanding of the frequency dependence of the resistance fluctuations in contacts, rectifying crystals, thin films, etc.

1. Introduction

WHEN a direct current is passed through a granular resistance such as a carbon microphone or a metallic-film grid leak, or through a single contact, there is produced a voltage fluctuation possessing a component called contact noise which is differentiated from the familiar thermal noise component by the fact that its r.m.s. value in any frequency band is roughly proportional to the magnitude of the average applied voltage, and is differentiated from shot noise by the strong frequency dependence of its spectral density. One may regard this component of the voltage fluctuation as arising from the spontaneous resistance fluctuations of the element in question if one is willing to allow the resistance to have a slight voltage dependence. This effect has been the subject of numerous experimental investigations,[1–8] among which we mention in particular that of Christensen and Pearson[9] on granular resistance elements. These authors (henceforth abbreviated as CP) arrived at an empirical formula, to be discussed presently, connecting the contact noise power per unit frequency band with the applied voltage, the resistance, and the frequency for several types of granular resistance. Their measurements covered a range of frequency from 60 to 10,000 cps, and involved the variation of several other parameters, i.e., pressure. More recently, Wegel and Montgomery[10] have measured the noise power arising

[1] H. A. Frederick, *Bell Telephone Quarterly* **10**, 164 (1931).
[2] A. W. Hull and N. H. Williams, *Phys. Rev.* **25**, 173 (1925).
[3] R. Otto, *Hochfrequenztechnikund Elektroakustik* **45**, 187 (1935).
[4] G. W. Barnes. *Jour. Franklin Inst.* **219**, 100 (1935).
[5] Erwin Meyer and Heinz Thiede, *E. N. T.* **12**, 237 (1935).
[6] F. S. Goucher, *Jour. Franklin Inst.* **217**, 407 (1934). *Bell Sys. Tech. Jour.* **13**, 163 (1934).
[7] J. Bernamont, *Annales de Phys.*, **1937**, 71–140.
[8] M. Surdin, *R. G. E.*, **47**, 97–101 (1940).
[9] C. J. Christensen and G. L. Pearson, *Bell Sys. Tech. Jour.* **15**, 197–223 (1936).
[10] Private communication.

from single contacts and have obtained results in agreement with the CP empirical formula down to frequencies of the order of $10^{-1} - 10^{-2}$ cps.

Significant theoretical work upon this problem has not been attempted until recently. G. G. Macfarlane[11] has advanced a theory based upon a non-linear mechanism containing one degree of freedom which seems to be in agreement with the CP law. W. Miller[12] has worked out a general theory of noise in crystal rectifiers. His theory is linear, contains essentially an infinite number of degrees of freedom, and is equivalent in many respects to the theory discussed in this paper; however, he has not succeeded in obtaining agreement with the experimental data on crystal rectifiers (which satisfy approximately the CP law) for any of the specific models he used.

The purpose of this paper is the calculation of the spectral density of the fluctuations of the electrical resistance when it is linearly coupled to a diffusing medium (particles or heat), or, mathematically speaking, is equal to a linear function of the concentration deviations of this diffusing medium. This diffusing medium undergoes thermally excited fluctuations and thereby causes fluctuations in the resistance. The motive behind this investigation was the understanding of the frequency dependence of contact noise discussed in the following paragraphs, but at the present time it is apparent that this treatment in addition may apply to rectifying crystals, thin films, transistors, etc. The quantitative details of the coupling between the resistance and the diffusing medium are not considered here; in consequence of which, this work can hardly pretend to give a complete explanation of contact noise. However, important results are given concerning the relation between the spectral density of the resistance, on one hand, and the geometry of the coupling and the dimensionality of the diffusion field on the other.

Now let us consider the CP empirical formula in detail. Let $\bar{R}$ be the average resistance[13] of the contact (we will henceforth consider only contacts and will regard a granular resistance as a contact assemblage) and let $R_1(t)$ be the instantaneous deviation from the average. By theorems I-3 of Appendix I, we can express the m.s. value of R_1 as a sum of the m.s. values of R_1 in each frequency interval as follows:

$$\overline{R_1^2} = \int_0^\infty S(\omega)\, d\omega, \tag{1.1}$$

[11] G. G. Macfarlane, *Proc. Phys. Soc.* **59**, Pt. 3, 366–374 (1947).

[12] To be published.

[13] The resistance of a contact is composed of two parts: the "gap resistance" and the "spreading resistance." The term "gap resistance" is self-explanatory. The "spreading resistance" is the resistance involved in driving the electric current through the body of the contact material along paths converging near the area of lowest gap resistance. The measured contact resistance is the sum of these two parts. In some of the particular physical models considered in Section 5, $\bar{R}$ is taken to be the gap resistance necessitating ad hoc arguments relating gap resistance and total resistance.

where $S(\omega)$ is called the spectral density of R_1 and ω is the frequency in radians per second. Now in our notation the CP formula may be expressed

$$S(\omega) = KV^{a-2}\bar{R}^{b+2}/\omega, \tag{1.2}$$

where V is the applied d-c voltage across the contact, K is a constant depending upon the temperature and the nature of the contact, and a and b are constants having values of about 1.85 and 1.25 respectively. CP state that the constant K is equal to about 1.2×10^{-10} in the case of a single carbon contact at room temperature.

In this paper we will regard the nonvanishing of $a - 2$ as arising from a non-linear effect which should become negligible at a sufficiently low voltage, although this interpretation does not seem completely justified on the basis of the work of CP. Consequently we assume that $a \to 2$ as $V \to 0$ in such a way that $V^{a-2} \to 1$. This is in keeping with the idea that the resistance fluctuations are truly spontaneous—at least for small applied voltages.

Although Eq. (1.2) may represent the observations over a large range of frequency it must break down at very high and very low frequencies in order that the noise power be finite (or, in other words, in order that the integral (1.1) converge).

One has several clues to be considered in looking for an underlying mechanism of the resistance fluctuations. First of all, the mechanical action of the thermal vibrations in the solid electrodes of the contact seems to be unimportant because of the following reasons: (1) there are no resonance peaks in $S(\omega)$ at the lowest characteristic frequencies of mechanical vibration of the contact assembly; (2) $S(\omega)$ becomes very large far below the lowest characteristic frequency; and (3), according to CP, $\overline{R_1^2}$ is strictly proportional to V^2 when the fluctuations are produced by acoustic noise vibrating the contact, whereas $\overline{R_1^2}$ is proportional to V^{a-2}, $a \sim 1.85$, when the fluctuations arise from the dominant mechanism existing in the macroscopically unperturbed contact. One of the obvious mechanisms left is a diffusional mechanism. Such a mechanism does not violate any of the observations to date and, furthermore, possesses a sufficient density of long relaxation times to give large contributions to $S(\omega)$ near zero frequency.

Evidence that diffusion of atoms (or ions) can be important in modulating a current is provided by the "flicker effect" in which the emission of electrons from a heated cathode is caused to fluctuate by the fluctuations in concentration of an adsorbed layer. We might suppose that contact noise is a different manifestation of the basic mechanism involved in the flicker effect.

In view of these considerations it seems worthwhile to investigate in a

general way a large class of models involving resistance fluctuations arising from diffusional mechanisms. In the next section we propose a general mathematical model embracing a class of linear diffusional mechanisms. In Sections 3 and 4 the consequences of the general mathematical model are obtained by the "Fourier" and "Smoluchowski" methods, respectively, these alternative methods leading to identical results. In Section 5, the general results are specialized to several physical cases, some of which are introduced only for the purpose of providing some insight into the relations between the possible physical mechanisms and the resultant resistance fluctuations, and one of which along with its refinement is a successful[14] attempt to provide a theory of Eq. (1.2). Section 6. is a summary.

2. The General Mathematical Model

The physical models which we consider in this paper are concerned with the fluctuations of contact resistance arising from a diffusional process. We are consequently led to consider the following general mathematical model embracing a rather extensive class of the physical models as special cases: Let us consider the instantaneous contact resistance $R(t)$ to be related to the intensity $c(\mathbf{r}, t)$ of some diffusing quantity as follows[15]:

$$G(R(t)) = \int F(\mathbf{r}, c(\mathbf{r}, t))\, d\mathbf{r}, \tag{2.1}$$

where $\mathbf{r}$ is a vector in two or three dimensional space depending on whether the diffusion takes place on a surface or in a volume, and $d\mathbf{r}$ is correspondingly a differential area or volume. The intensity $c(\mathbf{r}, t)$ may be either a concentration (in the case of diffusion of material in two or three dimensions) or a temperature (in the case of heat flow in three dimensions). In writing Eq. (2.1) we have evidently assumed that the contact resistance $R(t)$ is independent of the applied voltage. Eq. (2.1) may of course allow a dependence on voltage through the quantity c; however, we will consider no processes involving a dependence of c on the voltage. These restrictions, strictly speaking, make the model applicable only in the limit of low applied voltages.

Before proceeding further let us limit the treatment to the case in which the deviations of R and c from their average values are sufficiently small for higher powers of these deviations to be neglected. Let

$$R(t) = \bar{R} + R_1(t), \tag{2.2}$$

$$c(\mathbf{r}, t) = \bar{c} + c_1(\mathbf{r}, t), \tag{2.3}$$

[14] That is, successful in so far as agreement with the form of Eq. (1.2) is concerned.

[15] A relation more general than $R(t) = SF(\mathbf{r}, c(\mathbf{r}, t))\, d\mathbf{r}$ is required as one can see from considering the special case of a total resistance composed of a parallel array of resistive elements.

where $\bar{R}$ and $\bar{c}$ are the average[16] values of $R(t)$ and $c(r, t)$ respectively. Evidently, $\overline{R_1(t)} = 0$, and $\overline{c_1(r, t)} = 0$. Introducing the expressions (2.2) and (2.3) into Eq. (2.1), expanding in terms of $c_1(r, t)$, and neglecting terms of the order of c_1^2, we get

$$R_1(t) = \int f(r)\, c_1(r, t)\, dr, \tag{2.4}$$

where

$$f(r) = \left[\frac{\partial F(r, c)}{\partial c}\right]_{c=\bar{c}} \Big/ \left[\frac{dG(R)}{dR}\right]_{R=\bar{R}}$$

The function $f(r)$ defines the linear coupling between R_1 and c_1 and depends upon the specific physical model used. The non-linear terms neglected in Eq. (2.4) may be of importance under some conditions; however, we will not consider them here. Nevertheless, non-linear effects in the behavior of c_1 itself are possibly important in determining the form of the power spectrum of $R_1(t)$ in the neighborhood of zero frequency.

3. The Fourier Series Method of Solution

In this section we consider the state of the diffusing system to be defined by the Fourier space-amplitudes $c_k(t)$ of $c_1(r, t)$. The time behavior of $c_k(t)$ will be described by an infinite set of ordinary differential equations containing random exciting forces according to the conventional theory of Brownian motion.[17] This method yields the spectral density of $R_1(t)$ directly.

Now the diffusion process is assumed to occur in a rectangular area $A_2 = L_1 \times L_2$ or in a rectangular parallelopiped of volume $A_3 = L_1 \times L_2 \times L_3$. In regions of the above types, if we apply periodic boundary conditions[18], $c_1(r_0 t)$ may be expanded in Fourier space-series as follows:

$$c_1(r, t) = \sum_k{}' c_k(t) e^{ik \cdot r} \tag{3.1}$$

where the components of k take the values

$$k_i = 2\pi n_i / L_i, \quad i = 1, \cdots, \nu, \tag{3.2}$$

in which n_i are integers and ν is the number of dimensions. The prime on the summation indicates that the term for $k = 0$ is to be omitted. This is required by the equivalence of the time and space averages of c_1 (true for A_ν sufficiently large) and by the vanishing of the time average of c_1 (by definition).

[16] The average values here may be considered as either time or ensemble averages but not space averages.

[17] See Wang and Uhlenbeck, *Rev. Mod. Phys.*, **17**, 323–342 (1945).

[18] If the final results are given by integrals over k- space they will be insensitive to the boundary conditions.

Before proceeding to the solution itself let us consider what it is that we wish to know about $c_k(t)$. Expanding the function $f(\mathbf{r})$ of Eq. (2.4) in a Fourier space-series in the region A_ν,

$$f(\mathbf{r}) = \sum_k f_k e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{3.3}$$

we can write Eq. (2.4) in the form

$$R_1(t) = A_\nu \sum_k{}' f_k^* c_k(t) \tag{3.4}$$

where f_k^* is the conjugate of f_k.
The spectral density $S(\omega)$ of $R_1(t)$ is then

$$S(\omega) = A_\nu^2 \sum_{kk'}{}' C_{kk'}(\omega) f_k^* f_{k'}, \tag{3.5}$$

where $C_{kk'}(\omega)$ is the spectral density matrix for the set $c_k(t)$ given by

$$C_{kk'}(\omega) = 2\pi \operatorname*{Lim}_{\tau\to\infty} \frac{1}{\tau} [c_k(\omega, \tau) c_{k'}^*(\omega, \tau) + c_k(-\omega, \tau) c_{k'}^*(-\omega, \tau)] \tag{3.6}$$

in which

$$c_k(\omega, \tau) = \frac{1}{2\pi} \int_{-\tau/2}^{+\tau/2} c_k(t) e^{-i\omega t}\, dt. \tag{3.7}$$

For a full discussion of spectral densities and spectral density matrices see Appendix I. Consequently our objective in this section is the calculation of the maxtix $C_{kk'}(\omega)$ defined by Eq. (3.6).

Now we assume that $c_1(\mathbf{r}, t)$ satisfies the diffusion equation

$$\frac{\partial}{\partial t} c_1(\mathbf{r}, t) = D\bar{\nabla}^2 c_1(\mathbf{r}, t) + g(\mathbf{r}, t) \tag{3.8}$$

where D is a constant, ∇^2 is the Laplacian operator in two or three dimensions, and where $g(\mathbf{r}, t)$ is a random source function, whose Fourier space-amplitudes $g_k(t)$ possess statistical properties to be discussed presently. The random source function g is required for exciting c_1 sufficiently to maintain the fluctuations given by equilibrium theory. In the case of material diffusion the random source function g may be discarded in favor of a random force term of the form $-D/\chi\ T\cdot\vec{\nabla}\cdot[f(\bar{c} + c_1)]$, where $\nabla\cdot f = divf$, χ is the Boltzmann constant, T is the temperature, and f is the random force; however, in the linear approximation these two procedures will give identical final results. In the case of heat flow it is understood that the diffusion constant is $D = K/\rho C$ where K is the thermal conductivity, ρ the density, and C the specific heat. Eq. (3.8) as written is valid only for D a constant and c_1 small.

Introducing the expansion (3.1) and the expansion

$$g(\mathbf{r}, t) \sum_{\mathbf{k}}{}' g_{\mathbf{k}}(t) e^{i\mathbf{k}\cdot\mathbf{r}} \tag{3.9}$$

into Eq. (3.8) we obtain the infinite set of ordinary differential equations

$$\frac{d}{dt} c_{\mathbf{k}}(t) = -Dk^2 c_{\mathbf{k}}(t) + g_{\mathbf{k}}(t), \tag{3.10}$$

$$k = |\mathbf{k}|,$$

describing the time behavior of the Fourier space-amplitudes $c_{\mathbf{k}}(t)$. The Fourier space-amplitudes $g_{\mathbf{k}}(t)$ are assumed to be random functions of t possessing a white (flat) spectral density matrix $C_{\mathbf{kk}}$, independent of frequency. Multiplying Eq. (3.10) by $\frac{1}{2\pi} e^{-i\omega t}$, integrating with respect to time from $-\frac{1}{2}\tau$ to $+\frac{1}{2}\tau$, and neglecting the transients at the end points of the τ-interval, we obtain

$$c_{\mathbf{k}}(\omega, \tau) = \frac{g_{\mathbf{k}}(\omega, \tau)}{i\omega + Dk^2} \tag{3.11}$$

where $c_{\mathbf{k}}(\omega, \tau)$ is given by Eq. (3.7) and $g_{\mathbf{k}}(\omega, \tau)$ is given by an analogous equation. Forming the spectral density matrices we get for the diagonal elements

$$C_{\mathbf{kk}'}(\omega) = \frac{G_{\mathbf{kk}}}{\omega^2 + D^2 k^4} \tag{3.12}$$

The matrix $G_{\mathbf{kk}'}$ can now be evaluated by the thermodynamic theory of fluctuations (See Appendix II). This theory gives

$$c_{\mathbf{k}}(t) c^*_{\mathbf{k}'}(t) = \frac{\chi \delta_{\mathbf{kk}'}}{A_\nu s''} \tag{3.13}$$

where

$$\delta_{\mathbf{kk}'} = \begin{cases} 1 \text{ if } \mathbf{k} = \mathbf{k}' \\ 0 \text{ otherwise,} \end{cases}$$

$$s'' = -\left\{\frac{\partial^2 s}{\partial c^2}\right\}_{c=\bar{c}} + \frac{1}{\bar{T}} \left\{\frac{\partial^2 e}{\partial c^2}\right\}_{c=\bar{c}} \tag{3.14}$$

s and e being the entropy and energy, respectively, per unit area or volume, $\bar{T}$ the average temperature, and χ the Boltzmann constant. In the case where c is the concentration of particles whose configurational energy is constant, $s'' = \chi/\bar{c}$. If c be the temperature T then $s'' = C/\bar{T}^2$ where C

is the heat capacity per unit area or volume. Now by a general theorem concerning spectral density matrices (see Appendix I) we have

$$\overline{c_{\boldsymbol{k}}(l)c^{*}_{\boldsymbol{k}'}(l)} = \int_0^\infty C_{\boldsymbol{k}\boldsymbol{k}'}(\omega)\, d\omega,$$

giving finally by combination with (3.12) and (3.13),

$$G_{\boldsymbol{k}\boldsymbol{k}'} = \frac{2}{\pi}\frac{\chi D k^2}{A_\nu s''}\delta_{\boldsymbol{k}\boldsymbol{k}'}, \tag{3.15}$$

and

$$C_{\boldsymbol{k}\boldsymbol{k}'}(\omega) = \frac{2}{\pi}\frac{\chi D}{A_\nu s''}\frac{k^2\delta_{\boldsymbol{k}\boldsymbol{k}'}}{\omega^2 + D^2k^4}. \tag{3.16}$$

The spectral density $S(\omega)$ of $R_1(l)$ then becomes

$$S(\omega) = \frac{2}{\pi}\frac{\chi A_\nu D}{s''}\sum_{\boldsymbol{k}}\frac{k^2|f_{\boldsymbol{k}}|^2}{\omega^2 + D^2k^4}. \tag{3.17}$$

If we are concerned with frequencies greater than a characteristic frequency

$$\omega_0 = 4\pi^2 D/L^2 \tag{3.18}$$

where L is the smallest of L_i, $i = 1, \cdots, \nu$, then the summation in (3.17) may be replaced by an integration giving

$$S(\omega) = 2^{\nu+1}\pi^{\nu-1}\frac{\chi D}{s''}\int\frac{|f(\boldsymbol{k})|^2 k^2\, d\boldsymbol{k}}{\omega^2 + D^2k^4} \tag{3.19}$$

where

$$f(\boldsymbol{k}) = \frac{1}{(2\pi)^\nu}\int_{A_\nu} f(\boldsymbol{r})e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\, d\boldsymbol{r}. \tag{3.20}$$

The integration in Eq. (3.19) is carried out over the entire ν-dimensional $\boldsymbol{k}$-space. If the range of the function $f(\boldsymbol{r})$ is sufficiently small compared with the region A_ν, or if we let A_ν become indefinitely large, then the integration in Eq. (3.20) may be extended to all of ν-dimensional $\boldsymbol{r}$-space.

It is perhaps revealing to rephrase Eqs. (3.17) and (3.19) in terms of distributions of relaxation times. In the theory of dielectrics we speak of the real part of the dielectric constant being equal to a series of terms summed over a distribution of relaxation times: $\Sigma_i a_i\tau_i/(1 + \tau_i^2\omega^2)$, if the distribution is discrete, or $\int_0^\infty a(\tau)\tau d\tau/(1 + \tau^2\omega^2)$, if the distribution is continuous. In the above, a_i is the weight for the relaxation time τ_i, and, in

the case of a continuous distribution, $a(\tau)d\tau$ is the weight for the relaxation times in the range $d\tau$ containing τ. In these terms Eq. (3.17) becomes

$$S(\omega) = \sum_{k} \frac{a_k \tau_k}{1 + \tau_k^2 \omega^2}, \tag{3.17a}$$

where

$$a_k = \frac{2}{\pi} \frac{\chi}{s''} |f_k|^2. \tag{3.17b}$$

Eq. (3.19) becomes

$$S(\omega) = \int_0^\infty \frac{a(\tau)\tau \, d\tau}{1 + \tau^2 \omega^2} \tag{3.19a}$$

where

$$a(\tau) = \frac{2^\nu \pi^{\nu-1} \chi}{s'' D^{\nu/2} \tau^{\nu/2+1}} \int |f(\boldsymbol{l}/\sqrt{D\tau})|^2 \, d\Omega_\nu \tag{3.19b}$$

in which $\boldsymbol{l}$ is the unit vector in the direction of $\boldsymbol{k}$, $d\Omega_\nu$ is the differential "solid" angle in the ν-dimensional $\boldsymbol{k}$-space, and the integration is over the total solid angle (2π in 2 dimensions, or 4π, in 3).

It is of interest to calculate the self-covariance $\overline{R_1(t)R_1(t+u)}$. In Appendix I, it is shown that the self-covariance above is related to the spectral density $S(\omega)$ as follows:

$$\overline{R_1(t)R_1(t+u)} = \int_0^\infty S(\omega) \cos u\omega \, d\omega. \tag{3.21}$$

Using $S(\omega)$ in the form (3.17), Eq. (3.21) gives

$$\overline{R_1(t)R_1(t+u)} = \chi A_\nu / s'' \cdot \sum_{k} |f_k|^2 e^{-Duk^2}, \tag{3.22}$$

$$u > 0;$$

whereas with $S(\omega)$ in the form (3.19) we get

$$\overline{R_1(t)R_1(t+u)} = (2\pi)^\nu \chi / s'' \int |f(\boldsymbol{k})|^2 e^{-Duk^2} \, d\boldsymbol{k}. \tag{3.23}$$

The method of the next section yields the self-covariance directly.

4. Smoluchowski Method of Solution

We call the procedure employed in this section the "Smoluchowski method" because it is based on an equation very closely analogous to the well-known Smoluchowski equation forming the basis of the theory of

Markoff processes.[19] We set out directly to calculate the self-covariance for $R_1(t)$ which, by Eq. (2.4), is given by

$$\overline{R_1(t)R_1(t+u)} = \iint f(\mathbf{r}')f(\mathbf{r})\overline{c_1(\mathbf{r}'\,t)c_1(\mathbf{r}',\,t+u)}\,d\mathbf{r}'\,d\mathbf{r}. \tag{4.1}$$

Thus the problem is now reduced to the calculation of $\overline{c_1(\mathbf{r}',\,t)c_1(\mathbf{r},\,t+u)}$.

The quantity $\overline{c_1(\mathbf{r}',\,t)c_1(\mathbf{r},\,t+u)}$ is calculated in two steps. First we find $\overline{c_1(\mathbf{r},\,t+u)}^{(t+u)}$, the average value of c_1 at the point $\mathbf{r}$ at the time $t+u$ with the restriction that c_1 is known at every point $\mathbf{r}'$ with certainty to be $c_1(\mathbf{r}',\,t)$ at the time t (assuming, of course, that $u > 0$). Then we find that the required self-covariance for c_1 is given by multiplying the above $\overline{c_1(\mathbf{r},\,t+u)}^{(t+u)}$ by $c_1(\mathbf{r}',\,t)$ and averaging over-all values of $c_1(\mathbf{r},\,t)$ at time t; thus:

$$\overline{c_1(\mathbf{r}',\,t)\overline{c_1(\mathbf{r},\,t+u)}^{(t+u)}}^{(t)} = \overline{c_1(\mathbf{r}',\,t)c_1(\mathbf{r},\,t+u)}. \tag{4.2}$$

Now we assume that $\overline{c_1(\mathbf{r},\,t+u)}^{(t+u)}$ is related to $c(\mathbf{r}',\,t)$ by an integral equation, analogous to the Smoluchowski equation, as follows:

$$\overline{c_1(\mathbf{r},\,t+u)}^{(t+u)} = \int \rho(|\,\mathbf{r}-\mathbf{r}'\,|,\,u)c_1(\mathbf{r}',\,t)\,d\mathbf{r}'. \tag{4.3}$$

In the case that c represents a concentration as in the diffusion of particles, $\rho(|\,\mathbf{r}-\mathbf{r}'\,|,\,u)\,d\mathbf{r}$ is the probability that a particle be in the ν-dimensional volume element $d\mathbf{r}$ at time $t+u$ when it is known with certainty to be at $\mathbf{r}'$ a time t. Now the number of particles in $d\mathbf{r}'$ at $\mathbf{r}'$ at time t is evidently $[\bar{c}+c_1(\mathbf{r}',\,t)]\,d\mathbf{r}'$; consequently, the probable number of particles in $d\mathbf{r}$ at time $t+u$ which were in $d\mathbf{r}'$ at time t is $\rho(|\,\mathbf{r}-\mathbf{r}'\,|,\,u)[c+c_1(\mathbf{r}',\,t)]\,d\mathbf{r}\,d\mathbf{r}'$. Integration over $d\mathbf{r}'$ gives the total probable number $(\bar{c}+\overline{c_1(\mathbf{r},\,t+u)}^{(t+u)})\,d\mathbf{r}$ of particles in $d\mathbf{r}$ equal to $\left(\int \rho(|\,\mathbf{r}-\mathbf{r}'\,|,\,u)[\bar{c}+c_1(\mathbf{r}',\,t)]\,d\mathbf{r}'\right)d\mathbf{r}$ which reduces to $\left(\bar{c}+\int \rho(|\,\mathbf{r}-\mathbf{r}'\,|,\,u)c_1(\mathbf{r}',\,t)\,d\mathbf{r}'\right)d\mathbf{r}$. Division by $d\mathbf{r}$ and subtraction of $\bar{c}$ from both sides of the equality leads directly to Eq. (4.3). For the case of heat flow in crystal lattices the above picture can be used approximately if one uses the concept of phonons.[20] For a diffusional process $\rho(|\,\mathbf{r}-\mathbf{r}'\,|,\,u)$ is the normalized singular solution of the diffusion equation[21]; thus

$$\rho(|\,\mathbf{r}-\mathbf{r}'\,|,\,u) = \frac{1}{(4\pi Du)^{\nu/2}}\exp\left[-|\,\mathbf{r}-\mathbf{r}'\,|^2/4Du\right] \tag{4.4}$$

[19] Loc cit.
[20] J. Weigle, *Experientia*, **1**, 99-103 (1945).
[21] Chandrasekhar, *Rev. Mod. Phys.* **15**, 1 (1943).

where ν, as previously defined, is the number of dimensions of the region in which the process occurs.

Combining Eqs. (4.2) and (4.3) we get

$$\overline{c_1(\mathbf{r}', t)c_1(\mathbf{r}, t + u)} = \int \overline{c_1(\mathbf{r}', t)c_1(\mathbf{r}'', t)}^{(t)} \rho(|\mathbf{r} - \mathbf{r}''|, u)\, d\mathbf{r}''. \quad (4.5)$$

Now using the fact that

$$\overline{c_1(\mathbf{r}', t)c_1(\mathbf{r}'', t)}^{(t)} = \overline{c_1(\mathbf{r}', t)c_1(\mathbf{r}'', t)} \quad (4.6)$$

and using the relation

$$\overline{c_1(\mathbf{r}', t)c_1(\mathbf{r}'', t)} = \frac{\chi}{s''} \delta(\mathbf{r}' - \mathbf{r}'') \quad (4.7)$$

proved in Appendix II, Eq. (4.5) reduces to

$$\overline{c_1(\mathbf{r}', t)c_1(\mathbf{r}, t + u)} = \frac{\chi}{s''} \rho(|\mathbf{r} - \mathbf{r}'|, u). \quad (4.8)$$

Introducing the expression (4.8) into Eq. (4.1) we obtain at once the desired result

$$\begin{aligned} \overline{R_1(t)R_1(t + u)} &= \frac{\chi}{s''} \iint f(\mathbf{r})f(\mathbf{r}')\rho(|\mathbf{r} - \mathbf{r}'|, u)\, d\mathbf{r}\, d\mathbf{r}' \\ &= \frac{\chi}{s''(4\pi Du)^{\nu/2}} \iint f(\mathbf{r})f(\mathbf{r}') \exp[-|\mathbf{r} - \mathbf{r}'|/4Du]\, d\mathbf{r}\, d\mathbf{r}'. \end{aligned} \quad (4.9)$$

For the sake of comparison with Eq. (3.23) it is necessary to write (4.9) in terms of the Fourier space-transforms of the pertinent quantities. We write

$$f(\mathbf{r}) = \int f(\mathbf{k})\, e^{i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{k}$$

where

$$f(\mathbf{k}) = \frac{1}{(2\pi)^\nu} \int f(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{r}.$$

Also, we write

$$\begin{aligned} \rho(|\mathbf{r} - \mathbf{r}'|, u) &= \frac{1}{(4\pi Du)^{\nu/2}} \exp[-|\mathbf{r} - \mathbf{r}'|^2/4Du] \\ &= \frac{1}{(2\pi)^\nu} \int \exp[-Duk^2 + i\mathbf{k}\cdot(\mathbf{r} - \mathbf{r}')]\, d\mathbf{k}. \end{aligned}$$

After introduction of these expressions into (4.9) a short calculation yields the result

$$\overline{R_1(t)R_1(t + u)} = (2\pi)^\nu \frac{\chi}{s''} \int |f(\mathbf{k})|^2 e^{-Duk^2}\, d\mathbf{k} \quad (4.10)$$

which is identical with Eq. (3.23) (provided that we let $A_\nu \to \infty$ in the latter). Thus the methods of approach used in Section 3 and in this Section are completely equivalent.

5. Special Physical Models

In the previous two Sections we have developed by two different methods the consequences of the general mathematical model discussed in Section 2. Here we apply the general results to some special physical cases. In this task we will be principally concerned with finding the form of the function $f(\mathbf{r})$ and establishing the number of dimensions ν of the diffusion field. The main objective here is to provide some orientation on what mechanisms are or are not reasonable and to find at least one mechanism leading to the observed spectral density (inversely proportional to the frequency).

a. *A General Class of Models.* Here we consider all at once mechanisms which can be adequately represented by having $f(\mathbf{r})$ a ν-dimensional Gaussian function of the form

$$f(\mathbf{r}) = \prod_{i=1}^{\nu} \frac{b_i}{(2\pi\Delta_i)^{1/2}} e^{-x_i^2/2\Delta_i}, \tag{5.1}$$

where $\Delta_i^{1/2}$ is the "width" of the function measured along the i-th coordinate x_i. This form of $f(\mathbf{r})$ can represent approximately several types of localization of the coupling between R_1 and c_1, as will be seen in the special examples later. Now if we work with $A_\nu = \infty$, we will then have to consider the Fourier space-transform of $f(\mathbf{r})$, which is readily shown to be

$$\left.\begin{aligned} f(\mathbf{k}) &= \frac{1}{(2\pi)^\nu} \int f(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{r} \\ &= \prod_{i=1}^{\nu} b_i/2\pi \cdot e^{-\Delta_i k_i^2/2} \end{aligned}\right\} \tag{5.2}$$

Inserting this result into Eq. (3.19) we obtain immediately

$$S(\omega) = \frac{2\chi D}{\pi s''} \left(\prod_{i=1}^{\nu} \frac{b_i^2}{2\pi}\right) \int \frac{\exp\left(-\sum_{i=1}^{\nu} \Delta_i k_i^2\right) k^2\, d\mathbf{k}}{\omega^2 + D^2 k^4} \tag{5.3}$$

$$k = \sum_{i=1}^{\nu} k_i^2.$$

Inserting this expression (5.2) into Eq. (3.23) gives

$$\overline{R_1(t)R_1(t+u)} = \frac{\chi}{s''} \prod_{i=1}^{\nu} \frac{b_i^2}{[4\pi(\Delta_i + Du)]^{1/2}}; \qquad u > 0 \tag{5.4}$$

In order that (5.3) give the observed $S(\omega) \propto 1/\omega$ as a result, the integral must reduce to something proportional to $\int k\, dk/(\omega^2 + D^2k^4)$. It is clearly impossible that any choice of ν and any set of Δ_i can achieve this result. Furthermore the self-covariance $\overline{R_1(t)R_1(t+u)}$ corresponding to the observed $S(\omega)$ should depend explicitly on the way $S(\omega)$ deviates from $1/\omega$ as ω goes to zero. The expression (5.4) is finite for all $u > 0$ and does not depend upon any cut-off phenomena in $S(\omega)$ at low frequencies. Therefore we can exclude any physical mechanisms belonging to the class considered here. However, since several mechanisms that have been proposed do fall into this class, we consider them below:

(*i*) *Schiff's Mechanism.* Schiff[22] considered tentatively that the fluctuations in contact resistance may be due to the variation in concentration of diffusing ions (atoms, or molecules) in a high resistance region bounded by parallel planes of very small separation. Schiff arrived at a noise spectrum proportional to $1/\omega$ but at the expense of disagreeing in a fundamental way with thermodynamic fluctuation theory. Here we will show what the correct consequences of this mechanism are.

Consider that the high resistance region is bounded on either side by planes parallel to the (x_1, x_2)-plane and that the thickness in the x_3-direction is very small. Now this is obviously a case of the general model just considered in which we take $\nu = 3$ and

$$\left.\begin{aligned} \Delta_{1,2} &\gg Du, \\ \Delta_3 &\ll Du, \end{aligned}\right\} \tag{5.5}$$

where $1/u$ is of the order of magnitude of the frequencies of interest. It is then a matter of algebraic manipulation to show that

$$\begin{aligned} S(\omega) &\simeq \frac{\chi D}{s''} \cdot \frac{b_1^2 b_2^2 l_3^2}{2\pi^3 \Delta_1^{1/2} \Delta_2^{1/2}} \cdot \int_0^\infty \frac{k_1^2\, dk_1}{\omega^2 + D^2 k_1^4} \\ &= \frac{\chi}{D^{1/2} s''} \cdot \frac{b_1^2 b_2^2 b_3^2}{2^{5/2} \pi^2 \Delta_1^{1/2} \Delta_2^{1/2}} \frac{1}{\omega^{1/2}} \end{aligned} \tag{5.6}$$

and

$$\overline{R_1(t)R_1(t+u)} = \frac{\chi}{D^{1/2} s''} \cdot \frac{b_1^2 b_2^2 b_3^2}{(4\pi)^{3/2} \Delta_1^{1/2} \Delta_2^{1/2}} \cdot \frac{1}{u^{1/2}} \quad u > 0. \tag{5.7}$$

Thus we see that Shiff's mechanism leads to a noise spectrum proportional to $1/\omega^{1/2}$, not $1/\omega$. The explanation of the singularity of the self-covariance (5.7) at $u = 0$ lies in the inequalities (5.5).

[22] L. I. Schiff, *BuShips Contract NObs-34144*, "Tech. Rpt. #3", (1946). Before the publishing of this paper, Schiff informed the writer that he has discarded this mechanism.

The above treatment could just as well be applied to the case in which the diffusing quantity is heat instead of ions.

(ii) Resistance of a Localized Contact Disturbed by a Diffusing Surface Layer. Here we consider the case of two conductors covered with diffusing surface layers. It is supposed that the conduction from one conductor to the other is distributed Gaussianly with a width $\Delta^{1/2}$. Finally, it is supposed that the conductivity through the above area varies with the surface concentration of the surface layer in that region. This situation is well represented by the above general model by taking $\nu = 2$, $\Delta_1 = \Delta_2 = \Delta$, and $b_1 = b_2 = b$.

The self-covariance is readily calculated with the result

$$\overline{R_1(t)R_1(t+u)} = \frac{\chi}{s''} \cdot \frac{b^4}{4\pi(\Delta + Du)} \tag{5.8}$$

The corresponding spectral density is

$$S(\omega) = \frac{1}{2\pi^2}\frac{\chi b^4}{s''}\int_0^\infty \frac{\cos u\omega\, du}{\Delta + Du} = \frac{1}{2\pi^2}\frac{\chi b^4}{s''D}\left[-\cos(\omega\Delta/D)Ci(\omega\Delta/D) + \sin(\omega\Delta/D)\left(\frac{\pi}{2} - Si(\omega\Delta/D)\right)\right] \tag{5.8a}$$

where $Ci(x)$ and $Si(x)$ are the cosine and sine integrals[23] respectively. When $\omega \ll D/\Delta$

$$S(\omega) \simeq -\frac{1}{2\pi^2}\frac{\chi b^4}{s''D}\log(\delta\omega\Delta/D), \tag{5.8b}$$

$$\delta = 0.5772,$$

and when $\omega \gg D/\Delta$

$$S(\omega) \simeq \frac{1}{2\pi^2}\frac{\chi b^4 D}{s''\Delta^2}\frac{1}{\omega^2}. \tag{5.8c}$$

Thus we see that this case does not lead to the experimental form of the spectral density. It must be remarked that here $S(\omega)$ is very sensitive to the form of the self-covariance for small u.

b. Contact between Relatively Large Areas of Rough Surfaces Covered with Diffusing Surface Layers. We consider this case in detail since it leads to results in agreement with experiment. Furthermore, the more detailed consideration of this case will illustrate more fully the use of the general mathematical model, which may be of use in studying other diffusional mechanisms should they be postulated at some future time. This mechanism does not fall into the class just considered.

[23] See Jahnke and Emde, "Tables of Functions," p. 3, Dover (1943).

Suppose that the contact in an idealized form consists of two rough surfaces close together. Let positions on the surfaces be measured with respect to a plane between the surfaces, which we will call the mid-plane. Let the coordinate system be oriented so that the x_1 and x_2 axes lie in the mid-plane. Furthermore let the region in the mid-plane corresponding to close proximity of the rough surfaces be a rectangular area $A_2 = L_1 \times L_2$. Now, for convenience, we describe positions on the mid-plane by a two dimensional vector $\boldsymbol{r} = (x_1, x_2)$, and henceforth it will be understood that all vector expressions refer to this two-dimensional space. Let the distance between the upper and lower surfaces at $\boldsymbol{r}$ be denoted by $h(\boldsymbol{r})$. The geometry of the above model is illustrated in Fig. 1.

Now suppose that each surface is covered by a diffusing absorbed layer, such that the sum of the concentrations on both surfaces is $c(\boldsymbol{r}, t)$ at the time t in the neighborhood of $\boldsymbol{r}$. Now consider the conduction of current between the surfaces. Let us assume that the conductance per unit area (of mid-plane) is a function of the separation h of the surfaces and the total concentration c of absorbate near the point in question, i.e., $F(h, c)$. The total conductance will be the sum of the conductances through each element of area: hence, the instantaneous resistance $R(t)$ at time t will be given by

$$1/R(t) = \int_{A_2} F(h(\boldsymbol{r}), c(\boldsymbol{r}, t))\, d\boldsymbol{r} \tag{5.9}$$

where $d\boldsymbol{r}$ is the differential area on the mid-plane and the integration extends over the rectangle $A_2 = L_1 \times L_2$. Behind the above statements lies the tacit assumption that the radii of curvature of the rough surfaces are generally considerably larger than the values of h. However, we will not explicitly concern ourselves with this implied restriction.

At this point it is expedient to imagine that we have an ensemble of contacts identical in all respects except for different variations of the separation $h(\boldsymbol{r})$. If we have any function of h, $\psi(h)$ say, which we wish to average with respect to the variations of h, we simply average the function over the above ensemble giving a result which we denote by $\overline{\psi(h)}^{(e)}$.

Now let us write

$$h(\boldsymbol{r}) = \bar{h}^{(e)} + h_1(\boldsymbol{r}), \tag{5.10}$$

and, as before,

$$\left.\begin{aligned} R(t) &= \bar{R} + R_1(t), \\ c(\boldsymbol{r}, t) &= \bar{c} + c_1(\boldsymbol{r}, t). \end{aligned}\right\} \tag{5.11}$$

We assume that the ensemble average $\bar{h}^{(e)}$ and the time averages $\bar{R}$ and $\bar{c}$ are constants independent of $\boldsymbol{r}$ and t. Let us also assume that the integrals

of $h_1(r)$ and $c_1(r, t)$ over A_2 vanish. Inserting (5.10) and (5.11) into (5 9) and expanding, we get

$$\left.\begin{aligned} &1/\bar{R} - R_1(t)/\bar{R}^2 + \cdots = A_2 F(\bar{h}^{(\circ)}, \bar{c}) \\ &+ \left(\frac{\partial^2 F}{\partial h \partial c}\right)^{\circ} \int_{A_2} h_1(r) c_1(r, t)\, dr + \tfrac{1}{2}\left(\frac{\partial^2 F}{\partial h^2}\right)^{\circ} \int_{A_2} h_1^2(r)\, dr + \cdots, \end{aligned}\right\} \tag{5.12}$$

where the super zero on the derivatives indicates that they are evaluated at $h = \bar{h}^{(\circ)}$ and $c = \bar{c}$. In accordance with previous approximations in this memorandum we neglect[24] terms of the order of c_1^2 and R_1^2. We also neglect terms of the order of h_1^2. After taking the time average of (5.12) and subtracting the result from (5.12) we get

$$\left.\begin{aligned} R_1(t) &= \int_{A_2} f(r) c_1(r, t)\, dr, \\ f(r) &= \alpha \bar{R}^2 h_1(r), \\ \alpha &= -\left(\frac{\partial^2 F}{\partial h \partial c}\right)^{\circ} \end{aligned}\right\} \tag{5.13}$$

Thus we now have a special case of our general mathematical model for the number of dimensions $\nu = 2$, provided that we assume that the total concentration c on both of the rough surfaces fluctuates in the same manner as the concentration of a single adsorbed layer confined to a plane rectangular surface. The spectral density $S(\omega)$ of $R_1(t)$ is then given by Eq. (3.17) which we repeat here

$$S(\omega) = \frac{2}{\pi} \cdot \frac{\chi A_2 D}{s''} \cdot \frac{\Sigma'}{k} \frac{k^2 \mid f_k \mid^2}{\omega^2 + D^2 k^4} \tag{5.14}$$

where $\boldsymbol{k}$ is a two-dimensional vector whose components take the values $k_i = 2\pi n_i / L_i$, $n_i = 0, \pm 1, \pm 2, \cdots$, and where f_k are the Fourier space-amplitudes of $f(r)$ given by

$$f_k = A_2^{-1} \int_{A_2} f(r) e^{-i\boldsymbol{k}\cdot r}\, dr.$$

It may be appropriate at this point to consider the quantity s'' in detail for this particular case. If the energy e per unit area is independent of c, we have $s'' = -\dfrac{\partial^2 s}{\partial c^2}$ evaluated at $c = \bar{c}$ where s is here the entropy of the absorbate per unit area. For the sake of illustration let us consider a single layer of absorbate in which the molecules are non-interacting. If c, the sur-

[24] We neglect these terms not because they are small compared with c_1 or $h_1 c_1$ but, because they are non-fluctuating (in time), are hence to be compared with $1/\bar{R}$.

face concentration of the absorbate, be measured in molecules (atoms, or ions) per unit area, then, for the ideal system above, it follows that $s = -\chi c \log c$ and finally that $s'' = \chi/\bar{c}$. However, in the mechanism discussed in this part we have a compound system consisting of two separate layers on the upper and lower surfaces respectively. Nevertheless, a detailed analysis reveals that with c equal to the sum of the concentrations of both layers we still have $s'' = \chi/\bar{c}$ even though s itself is no longer given by an expression the same as that above. In conclusion let us consider the factor χ/s'' in Eq. (5.14). This factor is under the above idealization simply equal to $\bar{c}$. That is, the spectral density $S(\omega)$ is directly proportional to the average concentration of absorbed molecules, meaning simply that each molecule makes its contribution to the resistance fluctuations independently of the others. Of course, in any real system this will not be quite true; however, the existence of interactions will be manifested only by making χ/s'' not equal to $\bar{c}$ in Eq. (5.14).

The results quoted thus far apply to a system with a given $h(\mathbf{r})$. Now we shall average the right-hand side of Eq. (5.14) over the ensemble of variations of $h(\mathbf{r})$, it being supposed that $S(\omega)$ itself on the left-hand side will be negligibly affected by this operation. This amounts to replacing $|f_{\mathbf{k}}|^2$ by $\overline{|f_{\mathbf{k}}|^2}^{(e)}$. We then have

$$|f_{\mathbf{k}}|^{2(e)} = \alpha^2 \bar{R}^4 \overline{|h_{\mathbf{k}}|^2}^{(e)} \tag{5.15}$$

where $h_{\mathbf{k}}$ are the Fourier space-amplitudes of $h_1(\mathbf{r})$.

We now consider more closely the problem of calculating $\overline{|h_{\mathbf{k}}|^2}^{(e)}$. We want to assume that $h_1(\mathbf{r})$ is a more or less random function of $\mathbf{r}$. If $h_1(\mathbf{r})$ were a random function of $\mathbf{r}$ in the same way that the thermal noise voltage is a random function of the time t, then $\overline{|h_{\mathbf{k}}|^2}^{(e)}$ would be a constant independent of $\mathbf{k}$ and the self-covariance $\overline{h_1(\mathbf{r})h_1(\mathbf{r}')}^{(e)}$ would vanish for $\mathbf{r} \neq \mathbf{r}'$. This clearly cannot be so, since the function $h_1(\mathbf{r})$ with such statistical properties would represent a highly discontinuous type of surface incapable of physical existence. We then fall back upon the more reasonable assumption that the *gradient* of h_1 possesses statistical properties of the above type. This notion is precisely formulated by means of the following equations:

$$\vec{\nabla} h_1(\mathbf{r}) = \mathbf{p}(\mathbf{r}) \tag{5.16}$$

where

$$\int_{A_2} \mathbf{p}(\mathbf{r})\, d\mathbf{r} = 0, \tag{5.17}$$

and

$$\overline{\mathbf{p}(\mathbf{r})\mathbf{p}(\mathbf{r}')}^{(e)} = \beta \underset{\sim}{1} \delta(\mathbf{r} - \mathbf{r}'). \tag{5.18}$$

In Eq. (5.18) β is a parameter (with the dimensions of area) characterizing the amplitude of the surface roughness, and $\underset{\sim}{1}$ is the unit tensor in two dimensions. Expressing (5.16) in terms of the Fourier space-amplitudes h_k and p_k of h_1 and p respectively, we have

$$-ik\, h_k = p_k\,, \tag{5.19}$$

giving finally

$$\overline{|h_k|^2}^{(e)} = k \cdot \overline{p_k p_k^*}^{(e)} \cdot k/k^4 \tag{5.20}$$

Expressing (5.18) in terms of Fourier space-amplitudes we get

$$\overline{p_k p_{k'}^*}^{(e)} = \beta A_2^{-1} \underset{\sim}{1} \delta_{kk'}, \tag{5.21}$$

which, when inserted into (5.20) gives the following desired result:

$$\overline{|h_k|^2}^{(e)} = \beta A_2^{-1} k^{-2}. \tag{5.22}$$

Now replacing $|f_k|^2$ by $\overline{|f_k|^2}^{(e)}$ in Eq. (5.14) and substituting the expression (5.22) with the use of Eq. (5.15), we obtain

$$S(\omega) = \frac{2}{\pi} \cdot \frac{\chi D}{s''} \cdot \beta \alpha^2 \bar{R}^4 \sum_k \frac{1}{\omega^2 + D^2 k^4} \tag{5.23}$$

If the frequencies of interest are larger than a certain characteristic frequency $\omega_0 = 4\pi^2 D/L^2$ where L is the smaller of L_1 and L_2, the summation in (5.23) may be replaced by an integration giving finally

$$\left.\begin{aligned} S(\omega) &= \chi D/\pi^2 s'' \cdot \alpha^2 \beta \bar{R}^4 A_2 \cdot \int_0^\infty \frac{k\, dk}{\omega^2 + D^2 k^4} \\ &= \chi/4\pi s'' \cdot \beta \alpha^2 \bar{R}^4 A_2 \cdot 1/\omega \end{aligned}\right\} \tag{5.24}$$

This result is in agreement with experiment in most respects. The dependence on frequency is, of course, that experimentally observed by all investigators. The non-dependence on the voltage applied across the contact is implied by the basic assumptions common to all of the mechanisms considered here, and is in approximate agreement with the results of Christensen and Pearson (see Eq. (1.2)). For our result to agree with the results of CP as regards the dependence on the average resistance[25] $\bar{R}$, the factor $c^2 \bar{R}^4 A_2$ must be proportional to $\bar{R}^{2+b}$ where $b \sim 1.25$. These authors also imply that some of

[25] It must be remembered that the resistance $\bar{R}$ in the CP formula is the total contact resistance equal to sum of the gap resistance and the spreading resistance, whereas the $\bar{R}$ in our theory evidently should be considered the gap resistance. For the purposes of comparison we make the ad hoc assumption that the gap resistance is proportional to the total contact resistance.

the parameters necessary to complete the description of a contact between given substances at a given temperature show up implicitly only through $\bar{R}$. According to our theory the factor $\alpha^2 \bar{R}^4 A_2$ does not depend in any unique way upon $\bar{R}$; it matters by what means $\bar{R}$ is varied. If the resistance $\bar{R}$ is changed by altering the contact area A_2, keeping other parameters fixed, we would find that $\bar{R}A_2$ is constant so that the factor in question would be proportional to $\bar{R}^3$, that is, $b = 1$. However, if $\bar{R}$ is changed by varying the contact pressure, the effect would show up through the factor α^2, (β also, to some extent, perhaps) and, since one would expect α to increase somewhat with pressure whereas $\bar{R}$ decreases with pressure, the factor of interest would probably depend upon some power of $\bar{R}$ between 3 and 4, that is, $1 < b < 2$.

The theory formulated here suffers from the difficulty that the integral of the power spectrum with respect to frequency is logarithmically divergent at 0 and ∞, that is

$$\int_{\omega_1}^{\omega_2} S(\omega)\, d\omega \doteq \int_{\omega_1}^{\omega_2} d\omega\, \omega = \log(\omega_2/\omega_1) \to \infty \text{ as } \omega_1 \to 0 \text{ and } \omega_2 \to \infty .$$

The divergence at ∞ does not bother us as much as the divergence at 0 since, with only a divergence at ∞, the self-covariance $\overline{R_1(t)R_1(t+u)}$ exists for all non-vanishing values of u; whereas, with a singularity at 0, the self-covariance does not exist for any value of u. For this reason we cannot consider the self-covariance here. In Part *c* of this Section we consider a possible way of removing the divergence at 0, and consequently, then, we are able to calculate the self-covariance for non-vanishing values of u.

c. Refinement of the Theory of Part b. Here we propose a simple modification of the model of Part b, removing the divergence of the integral of $S(\omega)$ at $\omega = 0$. The modification considered here, although it is one of several possibilities any one of which is sufficient for removing the divergence (See Section 6.), is perhaps the only one that is sufficiently simple to treat in a memorandum of this scope. The results of this section are thus intended to be only provisional and suggestive.

Let us reconsider the statistics of the function $h(\mathbf{r})$ giving the separation between the surfaces near a point $\mathbf{r}$ on the mid-plane. The distribution of h's considered in the last section is open to several criticisms: (1) it possesses no characteristic length parallel to the mid-plane; and (2) the self-covariance $\overline{h_1(\mathbf{r})h_1(\mathbf{r}')}^{(e)}$ does not exist for any value of $\mathbf{r} - \mathbf{r}'$.

To correct partially for these difficulties we replace Eq. (5.22) by

$$\overline{|h_{\mathbf{k}}|^2}^{(e)} = \frac{\beta \ell^2}{A_2(1 + \ell^2 k^2)}, \tag{5.25}$$

where ℓ is a new characteristic length. The self-covariance $\overline{h_1(\mathbf{r})h_1(\mathbf{r}')}^{(e)}$ based upon (5.25) now exists for all values of $\mathbf{r} - \mathbf{r}'$ except 0. Thus we still

have the objection that the variance $\overline{h_1^{2(e)}}$ is infinite; however, this will cause us no trouble.

With Eq. (5.25) instead of (5.22) the spectral of density R_1 takes the form

$$\left.\begin{aligned} S(\omega) &= \chi D/\pi^2 s'' \cdot \beta\alpha^2 \bar{R}^4 A_2 \cdot \int_0^\infty \frac{\ell^2 k^2}{1 + \ell^2 k^2} \cdot \frac{k\, dk}{\omega^2 + D^2 k^4} \\ &= (\chi/4\pi s'') \cdot \beta\alpha^2 \bar{R}^4 A_2 \cdot 1/\omega \cdot Q(y), \\ Q(y) &= \frac{y\left(y - \frac{2}{\pi}\log y\right)}{1 + y^2}, \qquad y = \ell^2\omega/D. \end{aligned}\right\} \tag{5.26}$$

In obtaining the above equation we have made the usual assumption that the frequencies of interest are larger than $\omega_0 = 4\pi^2 D/L^2$, and have replaced the original sum by an integral. The function $Q(y)$ has the following properties:

$$\left.\begin{aligned} Q(y) &\simeq -\frac{2}{\pi} y \log y \quad \text{for} \quad y \ll 1 \\ Q(y) &\simeq 1 \quad \text{for} \quad y \gg 1 \end{aligned}\right\} \tag{5.27}$$

Hence for $\omega \ll D/\ell^2$, $S(\omega) \propto \log \omega$, the integral of which converges as $\omega \to 0$; whereas, for $\omega \gg D/\ell^2$, $S(\omega)$ differs negligibly from that given by the unrefined theory (Eq. (5.24)).

The self-covariance $\overline{R_1(t)\, R_1(t+u)}$ now exists for all non-vanishing u and is given by

$$\left.\begin{aligned} \overline{R_1(t)R_1(t+u)} &= (\chi/2\pi s'') \cdot \beta\alpha^2 \bar{R}^4 A_2 \cdot \int_0^\infty \frac{\ell^2 e^{-Duk^2}k\,dk}{1 + \ell^2 k^2} \\ &= (\chi/4\pi s'') \cdot \beta\alpha^2 \bar{R}^4 A_2 \cdot e^{Du/\ell^2}[-Ei(-Du/\ell^2)] \end{aligned}\right\} \tag{5.28}$$

where

$$\begin{aligned} -Ei(-x) &= \int_x^\infty e^{-v}\, dv/v, \\ &\simeq -\log \gamma x \quad \text{for} \quad x \ll 1, \\ &\simeq \frac{e^{-x}}{x} \quad \text{for} \quad x \gg 1, \\ \gamma &= 0.5772. \end{aligned}$$

Thus for $u \ll \ell^2/D$, $\overline{R_1(t)R_1(t+u)} \alpha - \log(\gamma Du/\ell^2)$ and for $u \gg \ell^2/D$, $\overline{R_1(t)R_1(t+u)} \alpha\ 1/u$.

Thus we have illustrated how one modification of the model has removed the divergence at $\omega = 0$.

It appears from the treatment here and in part b that roughness and diffusion in two dimensions are essential (at least in a linear treatment) features in obtaining $S(\omega)\alpha\ 1/\omega$. In the case of a non-linear coupling (to be considered in a later paper) a "self-induced" roughness effect may occur without introducing roughness ab initio as an intrinsic feature of the model.

6. Summary

(a) If the resistance deviation $R_1(t)$ is related to the concentration deviation $c_1(\boldsymbol{r}, t)$ of a diffusing medium (particles or heat) by the *linear* functional

$$R_1(t) = \int_{A_\nu} f(\boldsymbol{r}) c_1(\boldsymbol{r}, t)\, d\boldsymbol{r}, \tag{6.1}$$

where $\boldsymbol{r}$ is a vector and $d\boldsymbol{r}$ a volume element in a ν-dimensional space of volume A_ν, then the spectral density $S(\omega)$ of $R_1(t)$ is

$$S(\omega) = \frac{2}{\pi} \frac{\chi A_\nu D}{s''} \frac{k^2 |f_{\boldsymbol{k}}|^2}{\omega^2 + D^2 k^4}, \tag{6.2}$$

where D is the diffusion constant, s'' is defined by Eq. (3.14), χ is the Boltzmann constant, ω is the frequency (in radians per sec.), $\boldsymbol{k}$ is the wave number vector in ν-dimensional $\boldsymbol{k}$-space limited to a discrete lattice of points (defined by Eq. (3.2)) over which the summation is taken, and $f_{\boldsymbol{k}}$ is the $\boldsymbol{k}$th Fourier component of $f(\boldsymbol{r})$ (Eq. (3.3)).

(a) If the important terms in (6.2) vary slowly between lattice points in $\boldsymbol{k}$-space (true if $\omega > \omega_0$ given by Eq. (3.18)), then (6.1) can be replaced by the integral

$$S(\omega) = 2^{\nu+1} \pi^{\nu-1} \frac{\chi D}{s''} \int \frac{|f(\boldsymbol{k})|^2 k^2\, d\boldsymbol{k}}{\omega^2 + D^2 k^4}, \tag{6.3}$$

where the integration extends over the entire $\boldsymbol{k}$-space and where $f(\boldsymbol{k})$ is given by (Eq. 3.20)

$$f(\boldsymbol{k}) = \frac{1}{(2\pi)^\nu} \int_{A_\nu} f(\boldsymbol{r}) e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\, d\boldsymbol{r}. \tag{6.4}$$

(b) Let ω' be a frequency in the middle of a wide range. Suppose $|f(\boldsymbol{k})|^2$ averaged over the total solid angle in ν-dimensional $\boldsymbol{k}$-space is proportional to k^{2n}, where n is an integer, in a wide range of k with $k = \sqrt{\omega'/D}$ in its middle. It follows then that $S(\omega) \propto D^{-n-\nu/2}\, \omega^{-1+n+\nu/2}$ as long as $-1 < 2n + \nu + 1 < 3$. As a consequence, we see that with n an integer (as is true for the simple cases considered in Section 5) ν must be 2—the only even di-

mensionality—in order that $S(\omega)$ be inversely proportional to ω in agreement with experiment. In this case the only allowed value of n is -1.

(c) From (b) we have the interesting result that $S(\omega)$ is independent of D when it is inversely proportional to ω. This means that very slowly diffusing substances can contribute as much to contact noise as rapidly diffusing substances. This result can be derived on quite dimensional grounds and is not dependent upon the special assumptions underlying our treatment.

(d) A system comprising a high resistance layer modulated by the three-dimensional diffusion of particles or heat gives $S(\omega) \propto \omega^{-1/2}$. See Case a.(*i*) in Section 5.

(e) In a system composed of a localized contact disturbed by a diffusing surface layer (See Case a.(ii), Section 5), the self-covariance $\overline{R_1(t)R_1(t+u)}$ is inversely proportional to $\Delta + Du$ where Δ may be considered the contact area. We have $S(\omega) \propto -\log a + \text{const.}$ for $\omega \ll D/\Delta$ and $S(\omega) \propto \omega^{-2}$ for $\omega \gg D/\Delta$.

(f) In a system involving the contact between relatively large areas of rough surfaces covered with diffusing surface layers (Cases b. and c., Section 5), we have been successful in obtaining $S(\omega) \propto \omega^{-1}$, and also in obtaining a reasonable dependence upon the average resistance.

APPENDIX I

Spectral Density and the Self-Covariance

Here we consider in detail the spectral density, the self-covariance, and the relation between these two quantities, first for the case of a single random variable. The treatment is subsequently extended to the case of a *set* of random variables which necessitates the consideration of the spectral density matrix and the covariance matrix.

Let $y(t)$be a real random variable whose time average vanishes, $\overline{y(t)} = 0$. Now the m.s. value of y can be defined

$$\overline{y^2(t)} = \operatorname*{Lim}_{\tau\to\infty} \frac{1}{\tau} \int_{-\infty}^{+\infty} y^2(t, \tau)\, dt \tag{I-1}$$

where $y(t, \tau) = y(t)$ in the interval $-\frac{\tau}{2} \leq t < \frac{\tau}{2}$ and vanishes outside this interval. Evidently $y(t, \tau)$ can be expressed by the Fourier integral

$$y(t, \tau) = \int_{-\infty}^{+\infty} z(\omega, \tau) e^{i\omega t}\, d\omega \tag{I-2}$$

where

$$z(\omega, \tau) = \frac{1}{2\pi} \int_{-(\tau/2)}^{+(\tau/2)} y(t) e^{-i\omega t}\, dt.$$

By Parseval's theorem we obtain

$$\int_{-\infty}^{+\infty} y^2(t, \tau)\, dt = 2\pi \int_{-\infty}^{+\infty} |\, y(\omega, \tau)\,|^2\, d\omega,$$

which, when combined with (I-1), gives finally the desired result (using the fact that $|\, y(\omega, \tau)\,|^2$ is an even function of ω)

$$\overline{y^2(t)} = \int_0^\infty \Gamma(\omega)\, d\omega \tag{I-3}$$

where

$$\Gamma(\omega) = 4\pi \operatorname*{Lim}_{\tau\to\infty} \frac{1}{\tau} |\, y(\omega, \tau)\,|^2 \tag{I-4}$$

is the spectral density.

By a procedure not very different from the preceding, one can show that

$$\overline{y(t)y(t + u)} = \int_0^\infty \Gamma(\omega) \cos \omega u\, d\omega, \tag{I-5}$$

$$\Gamma(\omega) = \frac{2}{\pi} \int \overline{y(t)y(t + u)} \cos \omega u\, du. \tag{I-6}$$

The quantity $\overline{y(t)y(t + u)}$ is called the self-covariance.

Now let us suppose that we have a set of random variables $y_i(t)$ which are in general complex and whose time averages vanish. We are then led to consider, instead of (I-3), relations of the form

$$\overline{y_i(t)y_j^*(t)} = \int_0^\infty \Gamma_{ij}(\omega)\, d\omega \tag{I-7}$$

where now

$$\Gamma_{ij}(\omega) = 2\pi \operatorname*{Lim}_{\tau\to\infty} \frac{1}{\tau} [y_i(\omega, \tau)y_j^*(\omega, \tau) + y_i(-\omega, \tau)y_j^*(-\omega, \tau)] \tag{I-8}$$

in which

$$y_i(\omega, \tau) = \frac{1}{2\pi} \int_{-(\tau/2)}^{+(\tau/2)} y_i(t)e^{i\omega t}\, dt.$$

Instead of self-covariances like $\overline{y(t)y(t + u)}$ we have to consider a covariance matrix of the form $\overline{y_i(t)y_j^*(t + u)}$. Since we shall not have occasion in this paper to consider the relation between the spectral density matrix and the covariance matrix we will not consider the derivation of the analogue of Eq. (I-5).

APPENDIX II

Thermodynamic Theory of Fluctuations

The value of the quantity $\overline{c_1(\mathbf{r}, t)c_1(\mathbf{r}^1, t)}$ or $\overline{(c_k(t)c_{k^1}^*(t))}$ is determined from equilibrium considerations. Before going into the above continuum problem let us first consider the problem for the case of a system described by a finite set of variables. More specifically let us suppose that the state of the system subject to certain restraints (i.e. fixed total mass and energy) is described by the set of variables $x_1, \cdots, x_n$. Let the equilibrium state be given by the values $x_1^0, \cdots, x_n^0$, and let

$$x_i = x_i^0 + \alpha_i . \tag{II-1}$$

If the system is constrained to constant average energy E, the entropy of the non-equilibrium state $S = S^0 + \Delta S$ will be less than S^0, the entropy of the equilibrium state, by an amount

$$\Delta S = -\frac{1}{2}\sum_{ij} S_{ij}\alpha_i\alpha_j , \tag{II-2}$$

where

$$S_{ij} = -\left(\frac{\partial^2 S}{\partial x_i\,\partial x_j}\right)_{x_k=x_k^0} + \frac{1}{T^0}\left(\frac{\partial^2 E}{\partial x_i\,\partial x_j}\right)_{x_k=x_k^0}$$

Obviously, ΔS must be the negative of a positive definite quadratic form, otherwise the equilibrium state would not be a state of maximum entropy. The probability distribution[26] for the α's is given by

$$P(\alpha_1, \cdots, \alpha_n) = Ne^{\Delta S/\chi} \tag{II-3}$$

where N is a normalization factor and χ is the Boltzmann constant. Averaging the products $\alpha_i\,\alpha_j$ we find that

$$\sum_j S_{ij}\overline{\alpha_j\alpha_k} = \chi\delta_{ik} . \tag{II-4}$$

Multiplying (II-4) by the arbitrary set γ_i and summing over i we get

$$\sum_{ij}\gamma_i S_{ij}\overline{\alpha_j\alpha_k} = \chi\gamma_k . \tag{II-5}$$

The generalization to a system described by a continuous set of variables is not difficult on the basis of (II-5). Now suppose that, in a ν-dimensional space A_ν, we have a system whose state at time t is defined by the continuous set of values of the variable $c(\mathbf{r}, t) = \bar{c} + c_1(\mathbf{r}, t)$; we have

$$\Delta S = -\frac{1}{2}\int_{A_\nu} s''\,c_1^2(\mathbf{r}, t)\,d\mathbf{r} \tag{II-6}$$

[26] H. B. G. Casimir, *Rev. Mod. Phys.* **17**, Nos. 1 and 3, 343–4 (1945).

where

$$s'' = -\left(\frac{\partial^2 s}{\partial c^2}\right)_{c=\bar{c}} + \frac{1}{T^0}\left(\frac{\partial^2 e}{\partial c^2}\right)_{c=\bar{c}}$$

when s and e are the entropy and energy, respectively, per unit volume (of ν-dimensional space). In calculating (II-6) it was assumed that

$$\int_{A_\nu} c_1(\mathbf{r}, t)\, d\mathbf{r} = 0,$$

expressing the fact that the system is closed. In order to put (II-6) into a form strictly analogous to (II-2) we write it

$$\Delta S = -\frac{1}{2}\int_{A_\nu}\int_{A_\nu} s''\,\delta(\mathbf{r} - \mathbf{r}')c_1(\mathbf{r}, t)c_1(\mathbf{r}', t)\, d\mathbf{r}\, d\mathbf{r}'. \tag{II-7}$$

We see that the equation analogous to Eq. (II-5) must be

$$\int_{A_\nu}\int_{A_\nu} \gamma(\mathbf{r}')s''\,\delta(\mathbf{r} - \mathbf{r}'')\overline{c_1(\mathbf{r}'', t)c_1(\mathbf{r}, t)}\, d\mathbf{r}\, d\mathbf{r}' = \chi\gamma(\mathbf{r}) \tag{II-8}$$

where $\gamma(\mathbf{r})$ is an arbitrary function. Integrating (II-8) with respect to $\mathbf{r}''$ and using the fact that the delta function is defined by

$$\int \gamma(\mathbf{r}')\delta(\mathbf{r}' - \mathbf{r})\, d\mathbf{r}' = \gamma(\mathbf{r})$$

we readily arrive at the result

$$\overline{c_1(\mathbf{r}, t)c_1(\mathbf{r}', t)} = \frac{\chi}{s''}\,\delta(\mathbf{r} - \mathbf{r}'). \tag{II-9}$$

Using the Fourier space-expansions of C_1 and $\delta(\mathbf{r})$

$$c_1(\mathbf{r}, t) = \sum_{\mathbf{k}}{}' c_{\mathbf{k}}(t)e^{i\mathbf{k}\cdot\mathbf{r}},$$

$$\delta(\mathbf{r}) = \frac{1}{A_\nu}\sum_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}},$$

in the region $A_\nu = L_1 \times \cdots \times L_\nu$ with $k_i = 2\pi n_i/L_i$, we can write (II-9) over into the equivalent expression:

$$\overline{c_{\mathbf{k}}(t)c^*_{\mathbf{k}}(t)} = \frac{\chi}{A_\nu s''}\,\delta_{\mathbf{k}\mathbf{k}'}, \tag{II-10}$$

where

$$\delta_{\mathbf{k}\mathbf{k}'} = \begin{cases} 1 & \text{if } \mathbf{k} = \mathbf{k}', \\ 0 & \text{otherwise.} \end{cases}$$

Abstracts of Technical Articles by Bell System Authors

Audio-Frequency Measurements.[1] † W. L. Black* and H. H. Scott. This paper indicates the theory involved in making measurements of gain, frequency response, distortion, and noise at audio frequencies, with particular emphasis on such measurements made on high-gain systems. There are also discussed techniques of measurement and factors affecting the accuracy of results. This subject is not new art but has not previously been published in correlated form, to the knowledge of the authors.

Growing Quartz Crystals.[2] † E. Buehler and A. C. Walker. The Bell Telephone Laboratories started an investigation of this subject in March 1946, based on information gleaned from several investigators who visited Germany after the war, particularly Mr. J. R. Townsend of these Laboratories, and Professor A. C. Swinnerton of Antioch College. After a relatively few experiments made with equipment similar to that used by Professor Richard Nacken in Germany, and with the process he described, it became apparent that Nacken had made substantial progress in the art of growing quartz at temperatures and pressures near the critical state of water, i.e., about 374°C, and 3,200 pounds per square inch. This report summarizes further progress that has been made in the Laboratories since March 1946.

Corrosion of Telephone Outside Plant Material.[3] † K. C. Compton and A. Mendizza. Problems resulting from corrosion in the telephone outside plant are many and varied. In this article an attempt is made to give a broad overall picture of these problems and the manner in which they are met and solved by the telephone plant engineer.

Magnetic Recording in Motion Picture Techniques.[4] John G. Frayne and Halley Wolfe. Development of magnetic recording at the Bell Telephone Laboratories is described with the application of such facilities to Western Electric recording and reproducing systems. A method of driving 35-mm. magnetic film with a flutter content not greater than 0.1 per cent is described, as is a multigap erasing head.

Semi-Conducting Properties in Oxide Cathodes.[5] † N. B. Hannay, D. MacNair, and A. H. White. It has been widely assumed, without ade-

[1] *Proc. I. R. E.*, v. 37, pp. 1108–1115, October 1949.
* Of Bell Tel. Labs.
[2] *Sci. Monthly*, v. 69, pp. 148–155, September 1949.
[3] *Corrosion*, v. 5, pp. 194–197, June 1949.
[4] *S. M. P. E. Jour.*, v. 53, pp. 217–234, September 1949.
[5] *Jour. Applied Physics*, v. 20, pp. 669–681, July 1949.
† A reprint of this article may be obtained by writing to the Editor of the Bell System Technical Journal.

quate experimental verification, that barium-strontium oxide, as used in the oxide cathode, is an excess electronic semi-conductor. Accordingly, the electrical conductivity of (Ba,Sr)O has been studied as a function of temperature before and after activation with methane, extensive precautions being taken to exclude spurious effects. The increase in conductivity obtained characterizes (Ba,Sr)O as a "reduction" semi-conductor, and hence very probably as an electronic semi-conductor whose conduction electrons arise from a stoichiometric excess of (Ba,Sr) atoms in solid solution.

A basic prediction of the semi-conductor theory has been tested quantitatively with the finding that the electrical conductivity and the thermionic emission of a (Ba,Sr)O cathode are directly proportional through three orders of magnitude of activation; well-defined chemical and electrical activation and deactivation procedures were used in obtaining this result. It may be concluded that activation represents an increase in the chemical potential of the electrons in the oxide, little or no change in the state of the surface occurring. It has also been found that deviations from the proportionality of conductivity and emission may be expected under conditions leading to inhomogeneity in the oxide, in agreement with the semi-conductor theory also.

Electron Microscope and Diffraction Study of Metal Crystal Textures by Means of Thin Sections.[6] † R. D. HEIDENREICH. Bethe's dynamical theory of electron diffraction in crystals is developed using the approximation of nearly free electrons and Brillouin zones.

The use of Brillouin zones in describing electron diffraction phenomena proves to be illuminating since the energy discontinuity at a zone boundary is a fundamental quantity determining the existence of a Bragg reflection. The perturbation of the energy levels at a corner of a Brillouin zone is briefly discussed and the manner in which forbidden reflections may arise at a corner pointed out. It is concluded that the kinematic theory is inadequate for interpreting electron images of crystalline films.

An electrolytic method for preparing thin metal sections for electron microscopy and diffraction is introduced and its application to the structure of cold-worked aluminum and an aluminum-copper alloy demonstrated. It is concluded that cold-worked aluminum initially consists of small, inhomogeneously strained and disoriented blocks about 200A in size. These blocks are not revealed by etching but would contribute to line broadening in conventional diffraction experiments. By means of a reorientation of the blocks through a nucleation and growth process, larger disoriented domains about 1–3μ in size found experimentally could be accounted for. It is sug-

[6] *Jour. Applied Physics*, v. 20, pp. 993–1010, October 1949.

† A reprint of this article may be obtained by writing to the Editor of the Bell System Technical Journal.

gested that such a nucleation and growth reorientation phenomenon is responsible for self-recovering in cold-worked metals.

The formation of $CuAl_2$ precipitate particles is demonstrated with both electron micrographs and diffraction patterns. A fine lamellar structure found in the quenched Al-4 per cent Cu alloy is at present unexplained.

Path-Length Microwave Lenses.[7]† WINSTON E. KOCK. Lens antennas for microwave applications are described which produce a focusing effect by physically increasing the path lengths, compared to free space, of radio waves passing through the lens. This is accomplished by means of baffle plates which extend parallel to the magnetic vector, and which are either tilted or bent into serpentine shape so as to force the waves to travel the longer-inclined or serpentine path. The three-dimensional contour of the plate array is shaped to correspond to a convex lens. The advantages over previous metallic lenses are: broader band performance, greater simplicity, and less severe tolerances.

Refracting Sound Waves.[8]† WINSTON E. KOCK and F. K. HARVEY. Structures are described which refract and focus sound waves. They are similar in principle to certain recently developed electromagnetic wave lenses in that they consist of arrays of obstacles which are small compared to the wave-length. These obstacles increase the effective density of the medium and thus effect a reduced propagation velocity of sound waves passing through the array. This reduced velocity is synonymous with refractive power so that lenses and prisms can be designed. When the obstacles approach a half wave-length in size, the refractive index varies with wave-length and prisms then cause a dispersion of the waves (sound spectrum analyzer). Path length delay type lenses for focusing sound waves are also described. A diverging lens is discussed which produces a more uniform angular distribution of high frequencies from a loud speaker.

Double-Stream Amplifiers.[9]† J. R. PIERCE. This paper presents expressions useful in evaluating the gain of a double-stream amplifier having thin concentric electron streams of different velocity and input and output gaps across which both streams pass.

Direct Voltage Performance Test for Capacitor Paper.[10]† H. A. SAUER and D. A. MCLEAN. Performance of capacitors on accelerated life test may vary over a wide range depending upon the capacitor paper used. Indeed, at present a life test appears to be the only practical means for evaluating

[7] *Proc. I. R. E.*, v. 37, pp. 852–855, August 1949.
[8] *Acous. Soc. Amer. Jour.*, v. 21, pp. 471–481, September 1949.
[9] *Proc. I. R. E.*, v. 37, pp. 980–985, September 1949.
[10] *Proc. I. R. E.*, v. 37, pp. 927–931, August 1949.
† A reprint of this article may be obtained by writing to the Editor of the Bell System Technical Journal.

capacitor paper, since, within the limits observed in commercial material, the chemical and physical tests usually made do not correlate with life. Lack of correlation is ascribed to obscure physical factors which have not yet been identified.

Generally, several weeks are required to evaluate a paper by life tests of the usual severity. Unfortunately, the duration of these tests is too long for quality control of paper.

The desire for a life test which requires no more than a day or two for evaluation led to the development of a rapid d-c. test. The philosophy of rapid life testing is based upon the experimental evidence that the process of deterioration under selected temperature and voltage conditions is principally of a chemical nature, and also upon the well-known fact that rates of chemical reaction increase exponentially with temperature.

Life tests on two-layer capacitors conducted at 130°C. provide an acceleration in deterioration many fold more than that obtained in the lower-temperature life tests, and correlate well with these tests.

Contributors to this Issue

Sidney Darlington, Harvard University, B.S. in Physics, 1928; Massachusetts Institute of Technology, B.S. in E.E., 1929; Columbia University, Ph.D. in Physics, 1940. Bell Telephone Laboratories, 1929–. Dr. Darlington has been engaged in research in applied mathematics, with emphasis on network theory.

Richard C. Eggleston, Ph.B., 1909 and M.F., 1910, Yale University; U. S. Forest Service, 1910–1917; Pennsylvania Railroad, 1917–1920; First Lieutenant, Engineering Div., Ordnance Dept., World War I, 1918–1919. American Telephone and Telegraph Company, 1920–1927; Bell Telephone Laboratories, 1927–. Mr. Eggleston has been engaged chiefly with problems relating to the strength of timber and with statistical investigations in the timber products field.

J. R. Pierce, B.S. in Electrical Engineering, California Institute of Technology, 1933; Ph.D., 1936. Bell Telephone Laboratories, 1936–. Engaged in study of vacuum tubes.

S. O. Rice, B.S. in Electrical Engineering, Oregon State College, 1929; California Institute of Technology, 1929–30, 1934–35. Bell Telephone Laboratories, 1930–. Mr. Rice has been concerned with various theoretical investigations relating to telephone transmission theory.

J. M. Richardson, B.S., California Institute of Technology, 1941; Ph.D., Cornell, 1944. Bell Telephone Laboratories, 1945–49. Dr. Richardson at these Laboratories had been mainly associated with studies of ferroelectric materials, noise contacts, and contact erosion. At present he is with the Bureau of Mines at Pittsburgh.

VOLUME XXIX APRIL, 1950 NO. 2

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

50¢ per copy

$1.50 per Year

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the

American Telephone and Telegraph Company

195 Broadway, New York 7, N. Y.

SUBSCRIPTIONS

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

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XXVI *April, 1950* *No. 2*

Error Detecting and Error Correcting Codes

By R. W. HAMMING

1. Introduction

THE author was led to the study given in this paper from a consideration of large scale computing machines in which a large number of operations must be performed without a single error in the end result. This problem of "doing things right" on a large scale is not essentially new; in a telephone central office, for example, a very large number of operations are performed while the errors leading to wrong numbers are kept well under control, though they have not been completely eliminated. This has been achieved, in part, through the use of self-checking circuits. The occasional failure that escapes routine checking is still detected by the customer and will, if it persists, result in customer complaint, while if it is transient it will produce only occasional wrong numbers. At the same time the rest of the central office functions satisfactorily. In a digital computer, on the other hand, a single failure usually means the complete failure, in the sense that if it is detected no more computing can be done until the failure is located and corrected, while if it escapes detection then it invalidates all subsequent operations of the machine. Put in other words, in a telephone central office there are a number of parallel paths which are more or less independent of each other; in a digital machine there is usually a single long path which passes through the same piece of equipment many, many times before the answer is obtained.

In transmitting information from one place to another digital machines use codes which are simply sets of symbols to which meanings or values are attached. Examples of codes which were designed to detect isolated errors are numerous; among them are the highly developed 2 out of 5 codes used extensively in common control switching systems and in the Bell Relay

Computers,[1] the 3 out of 7 code used for radio telegraphy,[2] and the word count sent at the end of telegrams.

In some situations self checking is not enough. For example, in the Model 5 Relay Computers built by Bell Telephone Laboratories for the Aberdeen Proving Grounds,[1] observations in the early period indicated about two or three relay failures per day in the 8900 relays of the two computers, representing about one failure per two to three million relay operations. The self-checking feature meant that these failures did not introduce undetected errors. Since the machines were run on an unattended basis over nights and week-ends, however, the errors meant that frequently the computations came to a halt although often the machines took up new problems. The present trend is toward electronic speeds in digital computers where the basic elements are somewhat more reliable per operation than relays. However, the incidence of isolated failures, even when detected, may seriously interfere with the normal use of such machines. Thus it appears desirable to examine the next step beyond error detection, namely error correction.

We shall assume that the transmitting equipment handles information in the binary form of a sequence of 0's and 1's. This assumption is made both for mathematical convenience and because the binary system is the natural form for representing the open and closed relays, flip-flop circuits, dots and dashes, and perforated tapes that are used in many forms of communication. Thus each code symbol will be represented by a sequence of 0's and 1's.

The codes used in this paper are called *systematic* codes. Systematic codes may be defined[3] as codes in which each code symbol has exactly n binary digits, where m digits are associated with the information while the other $k = n - m$ digits are used for error detection and correction. This produces a *redundancy* R defined as the ratio of the number of binary digits used to the minimum number necessary to convey the same information, that is,

$$R = n/m.$$

This serves to measure the efficiency of the code as far as the transmission of information is concerned, and is the only aspect of the problem discussed in any detail here. The redundancy may be said to lower the effective channel capacity for sending information.

The need for error correction having assumed importance only recently, very little is known about the economics of the matter. It is clear that in

[1] Franz Alt, "A Bell Telephone Laboratories' Computing Machine"—I, II. Mathematical Tables and Other Aids to Computation, Vol. 3, pp. 1–13 and 60–84, Jan. and Apr. 1948.

[2] S. Sparks, and R. G. Kreer, "Tape Relay System for Radio Telegraph Operation," *R.C.A. Review*, Vol. 8, pp. 393–426, (especially p. 417), 1947.

[3] In Section 7 this is shown to be equivalent to a much weaker appearing definition.

using such codes there will be extra equipment for encoding and correcting errors as well as the lowered effective channel capacity referred to above. Because of these considerations applications of these codes may be expected to occur first only under extreme conditions. Some typical situations seem to be:

a. unattended operation over long periods of time with the minimum of standby equipment.
b. extremely large and tightly interrelated systems where a single failure incapacitates the entire installation.
c. signaling in the presence of noise where it is either impossible or uneconomical to reduce the effect of the noise on the signal.

These situations are occurring more and more often. The first two are particularly true of large scale digital computing machines, while the third occurs, among other places, in "jamming" situations.

The principles for designing error detecting and correcting codes in the cases most likely to be applied first are given in this paper. Circuits for implementing these principles may be designed by the application of well-known techniques, but the problem is not discussed here. Part I of the paper shows how to construct special minimum redundancy codes in the following cases:

a. single error detecting codes
b. single error correcting codes
c. single error correcting plus double error detecting codes.

Part II discusses the general theory of such codes and proves that under the assumptions made the codes of Part I are the "best" possible.

PART I

SPECIAL CODES

2. Single Error Detecting Codes

We may construct a single error detecting code having n binary digits in the following manner: In the first $n - 1$ positions we put $n - 1$ digits of information. In the n-th position we place either 0 or 1, so that the entire n positions have an even number of 1's. This is clearly a single error detecting code since any single error in transmission would leave an odd number of 1's in a code symbol.

The redundancy of these codes is, since $m = n - 1$,

$$R = \frac{n}{n-1} = 1 + \frac{1}{n-1}.$$

It might appear that to gain a low redundancy we should let n become very large. However, by increasing n, the probability of at least one error in a

symbol increases; and the risk of a double error, which would pass undetected, also increases. For example, if $p \ll 1$ is the probability of any error, then for n so large as $1/p$, the probability of a correct symbol is approximately $1/e = 0.3679\ldots$, while a double error has probability $1/2e = 0.1839\ldots$.

The type of check used above to determine whether or not the symbol has any single error will be used throughout the paper and will be called a *parity check*. The above was an *even* parity check; had we used an odd number of 1's to determine the setting of the check position it would have been an odd parity check. Furthermore, a parity check need not always involve all the positions of the symbol but may be a check over selected positions only.

3. Single Error Correcting Codes

To construct a single error correcting code we first assign m of the n available positions as information positions. We shall regard the m as fixed, but the specific positions are left to a later determination. We next assign the k remaining positions as check positions. The values in these k positions are to be determined in the encoding process by even parity checks over selected information positions.

Let us imagine for the moment that we have received a code symbol, with or without an error. Let us apply the k parity checks, in order, and for each time the parity check assigns the value observed in its check position we write a 0, while for each time the assigned and observed values disagree we write a 1. When written from right to left in a line this sequence of k 0's and 1's (to be distinguished from the values assigned by the parity checks) may be regarded as a binary number and will be called the *checking number*. We shall require that this checking number give the position of any single error, with the zero value meaning no error in the symbol. Thus the check number must describe $m + k + 1$ different things, so that

$$2^k \geq m + k + 1$$

is a condition on k. Writing $n = m + k$ we find

$$2^m \leq \frac{2^n}{n+1}.$$

Using this inequality we may calculate Table I, which gives the maximum m for a given n, or, what is the same thing, the minimum n for a given m.

We now determine the positions over which each of the various parity checks is to be applied. The checking number is obtained digit by digit, from right to left, by applying the parity checks in order and writing down the corresponding 0 or 1 as the case may be. Since the checking number is

TABLE I

n	m	Corresponding k
1	0	1
2	0	2
3	1	2
4	1	3
5	2	3
6	3	3
7	4	3
8	4	4
9	5	4
10	6	4
11	7	4
12	8	4
13	9	4
14	10	4
15	11	4
16	11	5
	Etc.	

to give the position of any error in a code symbol, any position which has a 1 on the right of its binary representation must cause the first check to fail. Examining the binary form of the various integers we find

$$\begin{aligned} 1 &= 1 \\ 3 &= 11 \\ 5 &= 101 \\ 7 &= 111 \\ 9 &= 1001 \end{aligned}$$

Etc.

have a 1 on the extreme right. Thus the first parity check must use positions

$$1, 3, 5, 7, 9, \cdots .$$

In an exactly similar fashion we find that the second parity check must use those positions which have 1's for the second digit from the right of their binary representation,

$$\begin{aligned} 2 &= 10 \\ 3 &= 11 \\ 6 &= 110 \\ 7 &= 111 \\ 10 &= 1010 \\ 11 &= 1011 \end{aligned}$$

Etc.,

the third parity check

$$
\begin{aligned}
4 &= 100 \\
5 &= 101 \\
6 &= 110 \\
7 &= 111 \\
12 &= 1100 \\
13 &= 1101 \\
14 &= 1110 \\
15 &= 1111 \\
20 &= 10100 \\
&\text{Etc.}
\end{aligned}
$$

It remains to decide for each parity check which positions are to contain information and which the check. The choice of the positions 1, 2, 4, 8, $\cdots$ for check positions, as given in the following table, has the advantage of making the setting of the check positions independent of each other. All other positions are information positions. Thus we obtain Table II.

TABLE II

Check Number	Check Positions	Positions Checked
1	1	1, 3, 5, 7, 9, 11, 13, 15, 17,$\cdots$
2	2	2, 3, 6, 7, 10, 11, 14, 15, 18,$\cdots$
3	4	4, 5, 6, 7, 12, 13, 14, 15, 20,$\cdots$
4	8	8, 9, 10, 11, 12, 13, 14, 15, 24,$\cdots$
.	.	.
.	.	.
.	.	.

As an illustration of the above theory we apply it to the case of a seven-position code. From Table I we find for $n = 7$, $m = 4$ and $k = 3$. From Table II we find that the first parity check involves positions 1, 3, 5, 7 and is used to determine the value in the first position; the second parity check, positions 2, 3, 6, 7, and determines the value in the second position; and the third parity check, positions 4, 5, 6, 7, and determines the value in position four. This leaves positions 3, 5, 6, 7 as information positions. The results of writing down all possible binary numbers using positions 3, 5, 6, 7, and then calculating the values in the check positions 1, 2, 4, are shown in Table III.

Thus a seven-position single error correcting code admits of 16 code symbols. There are, of course, $2^7 - 16 = 112$ meaningless symbols. In some applications it may be desirable to drop the first symbol from the code to avoid the all zero combination as either a code symbol or a code symbol plus a single error, since this might be confused with no message. This would still leave 15 useful code symbols.

TABLE III

Position							Decimal Value of Symbol
1	2	3	4	5	6	7	
0	0	0	0	0	0	0	0
1	1	0	1	0	0	1	1
0	1	0	1	0	1	0	2
1	0	0	0	0	1	1	3
1	0	0	1	1	0	0	4
0	1	0	0	1	0	1	5
1	1	0	0	1	1	0	6
0	0	0	1	1	1	1	7
1	1	1	0	0	0	0	8
0	0	1	1	0	0	1	9
1	0	1	1	0	1	0	10
0	1	1	0	0	1	1	11
0	1	1	1	1	0	0	12
1	0	1	0	1	0	1	13
0	0	1	0	1	1	0	14
1	1	1	1	1	1	1	15

As an illustration of how this code "works" let us take the symbol 0 1 1 1 1 0 0 corresponding to the decimal value 12 and change the 1 in the fifth position to a 0. We now examine the new symbol

0 1 1 1 0 0 0

by the methods of this section to see how the error is located. From Table II the first parity check is over positions 1, 3, 5, 7 and predicts a 1 for the first position while we find a 0 there; hence we write a

1 .

The second parity check is over positions 2, 3, 6, 7, and predicts the second position correctly; hence we write a 0 to the left of the 1, obtaining

0 1 .

The third parity check is over positions 4, 5, 6, 7 and predicts wrongly; hence we write a 1 to the left of the 0 1, obtaining

1 0 1 .

This sequence of 0's and 1's regarded as a binary number is the number 5; hence the error is in the fifth position. The correct symbol is therefore obtained by changing the 0 in the fifth position to a 1.

4. Single Error Correcting Plus Double Error Detecting Codes

To construct a single error correcting plus double error detecting code we begin with a single error correcting code. To this code we add one more posi-

tion for checking all the previous positions, using an even parity check. To see the operation of this code we have to examine a number of cases:

1. No errors. All parity checks, including the last, are satisfied.
2. Single error. The last parity check fails in all such situations whether the error be in the information, the original check positions, or the last check position. The original checking number gives the position of the error, where now the zero value means the last check position.
3. Two errors. In all such situations the last parity check is satisfied, and the checking number indicates some kind of error.

As an illustration let us construct an eight-position code from the previous seven-position code. To do this we add an eighth position which is chosen so that there are an even number of 1's in the eight positions. Thus we add an eighth column to Table III which has:

TABLE IV

0
0
1
1

1
1
0
0

1
1
0
0

0
0
1
1

PART II

GENERAL THEORY

5. A GEOMETRICAL MODEL

When examining various problems connected with error detecting and correcting codes it is often convenient to introduce a geometric model. The model used here consists in identifying the various sequences of 0's and 1's which are the symbols of a code with vertices of a unit n-dimensional cube. The code points, labelled x, y, z, $\cdots$, form a subset of the set of all vertices of the cube.

Into this space of 2^n points we introduce a *distance*, or, as it is usually called, a *metric*, $D(x, y)$. The definition of the metric is based on the observation that a single error in a code point changes one coordinate, two errors, two coordinates, and in general d errors produce a difference in d coordinates.

Thus we define the distance $D(x, y)$ between two points x and y as the number of coordinates for which x and y are different. This is the same as the least number of edges which must be traversed in going from x to y. This distance function satisfies the usual three conditions for a metric, namely,

$$D(x, y) = 0 \quad \text{if and only if } x = y$$

$$D(x, y) = D(y, x) > 0 \quad \text{if } x \neq y$$

$$D(z, y) + D(y, z) \geq D(x, z) \text{ (triangle inequality).}$$

As an example we note that each of the following code points in the three-dimensional cube is two units away from the others,

0 0 1
0 1 0
1 0 0
1 1 1 .

To continue the geometric language, a sphere of radius r about a point x is defined as all points which are at a distance r from the point x. Thus, in the above example, the first three code points are on a sphere of radius 2 about the point (1, 1, 1). In fact, in this example any one code point may be chosen as the center and the other three will lie on the surface of a sphere of radius 2.

If all the code points are at a distance of at least 2 from each other, then it follows that any single error will carry a code point over to a point that is *not* a code point, and hence is a meaningless symbol. This in turn means that any single error is detectable. If the minimum distance between code points is at least three units then any single error will leave the point nearer to the correct code point than to any other code point, and this means that any single error will be correctable. This type of information is summarized in the following table:

TABLE V

Minimum Distance	Meaning
1	uniqueness
2	single error detection
3	single error correction
4	single error correction plus double error detection
5	double error correction
	Etc.

Conversely, it is evident that, if we are to effect the detection and correction listed, then all the distances between code points must equal or exceed the minimum distance listed. Thus the problem of finding suitable codes is

the same as that of finding subsets of points in the space which maintain at least the minimum distance condition. The special codes in sections 2, 3, and 4 were merely descriptions of how to choose a particular subset of points for minimum distances 2, 3, and 4 respectively.

It should perhaps be noted that, at a given minimum distance, some of the correctability may be exchanged for more detectability. For example, a subset with minimum distance 5 may be used for:

a. double error correction, (with, of course, double error detection).

b. single error correction plus triple error detection.

c. quadruple error detection.

Returning for the moment to the particular codes constructed in Part I we note that any interchanges of positions in a code do not change the code in any essential way. Neither does interchanging the 0's and 1's in any position, a process usually called complementing. This idea is made more precise in the following definition:

Definition. Two codes are said to be *equivalent* to each other if, by a finite number of the following operations, one can be transformed into the other:

1. The interchange of any two positions in the code symbols.

2. The complementing of the values in any position in the code symbols.

This is a formal equivalence relation ($\sim$) since $A \sim A$; $A \sim B$ implies $B \sim A$; and $A \sim B$, $B \sim C$ implies $A \sim C$. Thus we can reduce the study of a class of codes to the study of typical members of each equivalence class.

In terms of the geometric model, equivalence transformations amount to rotations and reflections of the unit cube.

6. Single Error Detecting Codes

The problem studied in this section is that of packing the maximum number of points in a unit n-dimensional cube such that no two points are closer than 2 units from each other. We shall show that, as in section 2, 2^{n-1} points can be so packed, and, further, that any such optimal packing is equivalent to that used in section 2.

To prove these statements we first observe that the vertices of the n-dimensional cube are composed of those of two $(n-1)$-dimensional cubes. Let A be the maximum number of points packed in the original cube. Then one of the two $(n-1)$-dimensional cubes has at least $A/2$ points. This cube being again decomposed into two lower dimensional cubes, we find that one of them has at least $A/2^2$ points. Continuing in this way we come to a two-dimensional cube having $A/2^{n-2}$ points. We now observe that a square can have at most two points separated by at least two units; hence the original n-dimensional cube had at most 2^{n-1} points not less than two units apart.

To prove the equivalence of any two optimal packings we note that, if the packing is optimal, then each of the two sub-cubes has half the points. Calling this the first coordinate we see that half the points have a 0 and half have a 1. The next subdivision will again divide these into two equal groups having 0's and 1's respectively. After $(n - 1)$ such stages we have, upon reordering the assigned values if there be any, exactly the first $n - 1$ positions of the code devised in section 2. To each sequence of the first $n - 1$ coordinates there exist $n - 1$ other sequences which differ from it by one coordinate. Once we fix the n-th coordinate of some one point, say the origin which has all 0's, then to maintain the known minimum distance of two units between code points the n-th coordinate is uniquely determined for all other code points. Thus the last coordinate is determined within a complementation so that any optimal code is equivalent to that given in section 2.

It is interesting to note that in these two proofs we have used only the assumption that the code symbols are all of length n.

7. Single Error Correcting Codes

It has probably been noted by the reader that, in the particular codes of Part I, a distinction was made between information and check positions, while, in the geometric model, there is no real distinction between the various coordinates. To bring the two treatments more in line with each other we redefine a *systematic* code as a code whose symbol lengths are all equal and

1. The positions checked are independent of the information contained in the symbol.
2. The checks are independent of each other.
3. We use parity checks.

This is equivalent to the earlier definition. To show this we form a matrix whose i-th row has 1's in the positions of the i-th parity check and 0's elsewhere. By assumption 1 the matrix is fixed and does not change from code symbol to code symbol. From 2 the rank of the matrix is k. This in turn means that the system can be solved for k of the positions expressed in terms of the other $n - k$ positions. Assumption 3 indicates that in this solving we use the arithmetic in which $1 + 1 = 0$.

There exist non-systematic codes, but so far none have been found which for a given n and minimum distance d have more code symbols than a systematic code. Section 9 gives an example of a non-systematic code.

Turning to the main problem of this section we find from Table V that a single error correcting code has code points at least three units from each other. Thus each point may be surrounded by a sphere of radius 1 with no two spheres having a point in common. Each sphere has a center point and

n points on its surface, a total of $n + 1$ points. Thus the space of 2^n points can have at most:

$$\frac{2^n}{n+1}$$

spheres. This is exactly the bound we found before in section 3.

While we have shown that the special single error correcting code constructed in section 3 is of minimum redundancy, we cannot show that all optimal codes are equivalent, since the following trivial example shows that this is not so. For $n = 4$ we find from Table I that $m = 1$ and $k = 3$. Thus there are at most two code symbols in a four-position code. The following two optimal codes are clearly not equivalent:

$$\begin{matrix} 0\ 0\ 0\ 0 \\ 1\ 1\ 1\ 1 \end{matrix} \quad \text{and} \quad \begin{matrix} 0\ 0\ 0\ 0 \\ 0\ 1\ 1\ 1 \end{matrix} .$$

8. Single Error Correcting Plus Double Error Detecting Codes

In this section we shall prove that the codes constructed in section 4 are of minimum redundancy. We have already shown in section 4 how, for a minimum redundancy code of $n - 1$ dimensions with a minimum distance of 3, we can construct an n dimensional code having the same number of code symbols but with a minimum distance of 4. If this were not of minimum redundancy there would exist a code having more code symbols but with the same n and the same minimum distance 4 between them. Taking this code we remove the last coordinate. This reduces the dimension from n to $n - 1$ and the minimum distance between code symbols by, at most, one unit, while leaving the number of code symbols the same. This contradicts the assumption that the code we began our construction with was of minimum reduncancy. Thus the codes of section 4 are of minimum redundancy.

This is a special case of the following general theorem: To any minimum redundancy code of N points in $n - 1$ dimensions and having a minimum distance of $2k - 1$ there corresponds a minimum redundancy code of N points in n dimensions having a minimum distance of $2k$, and conversely. To construct the n dimensional code from the $n - 1$ dimensional code we simply add a single n-th coordinate which is fixed by an even parity check over the n positions. This also increases the minimum distance by 1 for the following reason: Any two points which, in the $n - 1$ dimensional code, were at a distance $2k - 1$ from each other had an odd number of differences between their coordinates. Thus the parity check was set oppositely for the two points, increasing the distance between them to $2k$. The additional coordinate could not decrease any distances, so that all points in the code are now at a minimum distance of $2k$. To go in the reverse direction we simply

drop one coordinate from the n dimensional code. This reduces the minimum distance of $2k$ to $2k - 1$ while leaving N the same. It is clear that if one code is of minimum redundancy then the other is, too.

9. Miscellaneous Observations

For the next case, minimum distance of five units, one can surround each code point by a sphere of radius 2. Each sphere will contain

$$1 + C(n, 1) + C(n, 2)$$

points, where $C(n, k)$ is the binomial coefficient, so that an upper bound on the number of code points in a systematic code is

$$\frac{2^n}{1 + C(n, 1) + C(n, 2)} = \frac{2^{n+1}}{n^2 + n + 2} \geq 2^m.$$

This bound is too high. For example, in the case of $n = 7$, we find that $m = 2$ so that there should be a code with four code points. The maximum possible, as can be easily found by trial and error, is two.

In a similar fashion a bound on the number of code points may be found whenever the minimum distance between code points is an odd number. A bound on the even cases can then be found by use of the general theorem of the preceding section. These bounds are, in general, too high, as the above example shows.

If we write the bound on the number of code points in a unit cube of dimension n and with minimum distance d between them as $B(n, d)$, then the information of this type in the present paper may be summarized as follows:

$$B(n, 1) = 2^n$$

$$B(n, 2) = 2^{n-1}$$

$$B(n, 3) = 2^m \leq \frac{2^n}{n + 1}$$

$$B(n, 4) = 2^m \leq \frac{2^{n-1}}{n}$$

$$B(n - 1, 2k - 1) = B(n, 2k)$$

$$B(n, 2k - 1) = 2^m \leq \frac{2^n}{1 + C(n, 1) + \cdots + C(n, k - 1)}.$$

While these bounds have been attained for certain cases, no general methods have yet been found for contructing optimal codes when the minimum distance between code points exceeds four units, nor is it known whether the bound is or is not attainable by systematic codes.

We have dealt mainly with systematic codes. The existence of non-systematic codes is proved by the following example of a single error correcting code with $n = 6$.

$$\begin{array}{c} 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 1\ 0\ 1\ 0\ 1 \\ 1\ 0\ 0\ 1\ 1\ 0 \\ 1\ 1\ 1\ 0\ 0\ 0 \\ 0\ 0\ 1\ 0\ 1\ 1 \\ 1\ 1\ 1\ 1\ 1\ 1\ . \end{array}$$

The all 0 symbol indicates that any parity check must be an even one. The all 1 symbol indicates that each parity check must involve an even number of positions. A direct comparison indicates that since no two columns are the same the even parity checks must involve four or six positions. An examination of the second symbol, which has three 1's in it, indicates that no six-position parity check can exist. Trying now the four-position parity checks we find that

$$\begin{array}{c} 1\ 2\quad\ 5\ 6 \\ 2\ 3\ 4\ 5 \end{array}$$

are two independent parity checks and that no third one is independent of these two. Two parity checks can at most locate four positions, and, since there are six positions in the code, these two parity checks are not enough to locate any single error. The code is, however, single error correcting since it satisfies the minimum distance condition of three units.

The only previous work in the field of error correction that has appeared in print, so far as the author is aware, is that of M. J. E. Golay.[4]

[4] M. J. E. Golay, Correspondence, Notes on Digital Coding, *Proceedings of the I.R.E.*, Vol. 37, p. 657, June 1949.

Optical Properties and the Electro-optic and Photoelastic Effects in Crystals Expressed in Tensor Form

By W. P. MASON

I. Introduction

THE electro-optic and photoelastic effects in crystals were first investigated by Pöckels,[1] who developed a phenomenological theory for these effects and measured the constants for a number of crystals. Since then not much work has been done on the subject till the very large electro-optic effects were discovered in two tetragonal crystals ammonium dihydrogen phosphate (ADP) and potassium dihydrogen phosphate (KDP). With these crystals light modulators can be obtained which work on voltages of 2000 volts or less. Their use has been suggested[2] in such equipment as light valves for sound on film recording and in television systems. Furthermore, since the electro-optic effect depends on a change in the dielectric constant with voltage, and the dielectric constant is known to follow the field up to 10^{10} cycles, it is obvious that this effect can be used to produce very short light pulses which may be of interest for physical investigations and for stroboscopic instruments of very high resolution. Hence these crystals renew an interest in the electro-optic effect.

In looking over the literature on the electro-optic effect and photoelastic effect in crystals, there do not seem to be any derivations that give them in terms of thermodynamic potentials, which allow one to investigate the condition under which equalities occur between the various electro-optic and photoelastic constants. Hence it is the purpose of this paper to give such a derivation. Another object is to give a derivation of Maxwell's equations in tensor form, and to apply them to the derivation of the Fresnel ellipsoid.

The first sections deal with the optics of crystals, and derive the Fresnel ellipsoid from Maxwell's equations. Other sections give a derivation of the two effects, discuss methods for measuring them by determining the birefrigence in various directions and give the constants for the two effects in terms of crystal symmetries. The final section discusses the application of the photoelastic effect for measuring strains in isotropic media.

[1] F. Pöckels, Lehrbuck Der Kristalloptic, B. Teubner, Leipzig, 1906.

[2] See *Patent* 2,467,325 issued to the writer; "Light Modulation by P type Crystals," George D. Gotschall, *Jour. Soc. Motion Picture Engineers,* July, 1948, pp. 13–20; B. H. Billings, *Jour. Opt. Soc. Am.*, 39, 797, 802 (1949).

II. Solution of Maxwell's Equations In Tensor Form

In tensor notation, Maxwell's equations for a nonmagnetic medium with no free charges take the form

$$\frac{1}{V}\frac{\partial D_i}{\partial t} = \epsilon_{ijk}\frac{\partial H_j}{\partial x_k}; \quad \frac{1}{V}\frac{\partial H_j}{\partial t} = -\epsilon_{jki}\frac{\partial E_k}{\partial x_i}; \quad \frac{\partial D_i}{\partial x_i} = 0; \quad \frac{\partial H_j}{\partial x_j} = 0 \tag{1}$$

where D_i is the electric displacement, H_j the magnetic field, E_k the electric field, V the velocity of light in vacuo and ϵ_{ijk} a tensor equal to zero when $i = j$ or k or $j = k$, but equal to 1 or -1 when all three numbers are different. If the numbers are in rotation, i.e. 1, 2, 3; 2, 3, 1; 3, 1, 2 the value is $+1$ while, if they are out of rotation, the value is -1.

We assume the electric vector to be representable by a plane wave whose planes of equal phase are taken normal to the unit vector n_i. Then

$$E_k = E_{0_k} e^{j\omega(t - x_i n_i/v)} \tag{2}$$

where E_{0_k} are constants representing the maximum values of the field along the three rectangular coordinates and $j = \sqrt{-1}$. Substituting (2) in the second of equations (1), noting that E_{0_k} are not functions of the space coordinates, we have

$$\frac{1}{V}\frac{\partial H_j}{\partial t} = \frac{j\omega}{v}[\epsilon_{jki} E_{0_k} n_i] e^{j\omega[t - x_i n_i/v]}. \tag{3}$$

Integrating with respect to the time

$$H_j = \frac{V}{v}[\epsilon_{jki} E_{0_k} n_i] e^{j\omega[t - x_i n_i/v]} = H_{0_j} e^{j\omega(t - x_i n_i/v)}. \tag{4}$$

Hence,

$$H_{0_j} = \frac{V}{v}[\epsilon_{jki} E_{0_k} n_i] \tag{5}$$

and therefore the magnetic vector is normal to the plane determined by E_{0_k} and n_i.

Next, using the first of equations (1),

$$\begin{aligned}\frac{\partial D_i}{\partial t} = V\epsilon_{ijk}\frac{\partial H_j}{\partial x_k} = V\epsilon_{ijk} H_{0_j}\frac{\partial e^{j\omega[t - x_k n_k/v]}}{\partial x_k} \\ = -\frac{j\omega V}{v}[\epsilon_{ijk} H_{0_j} n_k] e^{j\omega[t - x_k n_k/v]}.\end{aligned} \tag{6}$$

Integrating with respect to time,

$$D_i = -\frac{V}{v}[\epsilon_{ijk} H_{0_j} n_k] e^{j\omega[t - x_k n_k/v]}. \tag{7}$$

Inserting the value of H_{0_j} from (5), this equation takes the form

$$D_i = -\frac{V^2}{v^2}\,[\epsilon_{ijk}(\epsilon_{jki}\,E_{0_k}\,n_i)n_k]e^{j\omega[t-x_i n_i/v]}$$

and, in general,

$$D_i = -\frac{V^2}{v^2}\,[\epsilon_{ijk}(\epsilon_{jki}\,E_k n_i)n_k]. \tag{9}$$

Expanding the inner parenthesis, we have the components

$$(E_2n_3 - E_3n_2)_1; \qquad (E_3n_1 - E_1n_3)_2; \qquad (E_1n_2 - E_2n_1)_3. \tag{10}$$

Then

$$\epsilon_{ijk}[(E_2n_3 - E_3n_2);\ (E_3n_1 - E_1n_3);\ (E_1n_2 - E_2n_1)]n_k \text{ gives}$$

$$\begin{aligned}
D_1 &= -\frac{V^2}{v^2}\,[(E_3\,n_1 - E_1\,n_3)n_3 - (E_1\,n_2 - E_2\,n_1)n_2] \\
&= [(E_3\,n_3 + E_2\,n_2 + E_1\,n_1)n_1 - E_1(n_1^2 + n_2^2 + n_3^2)] \\
D_2 &= -\frac{V^2}{v^2}\,[(E_1\,n_2 - E_2\,n_1)n_1 - (E_2\,n_3 - E_3\,n_2)n_3] \\
&= [(E_3\,n_3 + E_2\,n_2 + E_1\,n_1)n_2 - E_2(n_1^2 + n_2^2 + n_3^2)] \\
D_3 &= -\frac{V^2}{v^2}\,[(E_2\,n_3 - E_3\,n_2)n_2 - (E_3\,n_1 - E_1\,n_3)n_1] \\
&= [(E_3\,n_3 + E_2\,n_2 + E_1\,n_1)n_3 - E_3(n_1^2 + n_2^2 + n_3^2)].
\end{aligned} \tag{11}$$

Now, since $n_1^2 + n_2^2 + n_3^2 = 1$ because n is a unit vector, we have

$$D_i = \frac{V^2}{v^2}\,[E_i - (E_j\,n_j)n_i] \quad \text{or} \quad \frac{v^2}{V^2}\,D_i - E_i - (E_j\,n_j)n_i = 0. \tag{12}$$

This equation states that D_i, E_i and n_i are in the same plane, H_j being normal to the plane as shown by Fig. 1. The energy flow vector

$$S_i = \frac{V^2}{4\pi}\,\epsilon_{ijk}\,E_j\,H_k \tag{13}$$

also lies in the plane since it is perpendicular to E and H. It is at the same angle θ with n that E is with D. The velocity of energy flow is $v/\cos\theta$. The energy velocity is called the ray velocity and the energy path the ray path.

Next, from the relation for a material medium, that

$$D_i = K_{ij}E_j \text{ or conversely } E_j = \beta_{ji}D_i \tag{14}$$

where K_{ij} are the dielectric constants measured at optical frequencies and β_{ji} are the impermeability constants determined from the relations

$$\beta_{ji} = \Delta^{ji}/\Delta^{K} \tag{15}$$

where

$$\Delta^{K} = \begin{vmatrix} K_{11} & K_{12} & K_{13} \\ K_{12} & K_{22} & K_{23} \\ K_{13} & K_{23} & K_{33} \end{vmatrix}$$

and Δ^{ji} the determinant obtained by suppressing the j^{th} row and i^{th} column, we can eliminate E_i from equation (12) and obtain

$$\begin{aligned} \frac{v^2}{V^2} D_1 &= \beta_{11} D_1 + \beta_{12} D_2 + \beta_{13} D_3 - (E_j n_j) n_1 \\ \frac{v^2}{V^2} D_2 &= \beta_{12} D_1 + \beta_{22} D_2 + \beta_{23} D_3 - (E_j n_j) n_2 \\ \frac{v^2}{V^2} D_3 &= \beta_{13} D_1 + \beta_{23} D_2 + \beta_{33} D_3 - (E_j n_j) n_3. \end{aligned} \tag{16}$$

This can be put in the form

$$\begin{aligned} (E_j n_j) n_1 &= D_1[\beta_{11} - v^2/V^2] + \beta_{12} D_2 + \beta_{13} D_3 \\ (E_j n_j) n_2 &= \beta_{12} D_1 + (\beta_{22} - v^2/V^2) D_2 + \beta_{23} D_3 \\ (E_j n_j) n_3 &= \beta_{13} D_1 + \beta_{23} D_2 + (\beta_{33} - v^2/V^2) D_3. \end{aligned} \tag{17}$$

Solving for D_1, D_2 and D_3

$$\begin{aligned} D_1 &= [(\beta_{22} - v^2/V^2)(\beta_{33} - v^2/V^2) - \beta_{23}^2][E_j n_j] n_1 \\ D_2 &= [(\beta_{11} - v^2/V^2)(\beta_{33} - v^2/V^2) - \beta_{13}^2][E_j n_j] n_2 \\ D_3 &= [(\beta_{11} - v^2/V^2)(\beta_{22} - v^2/V^2) - \beta_{12}^2][E_j n_j] n_3. \end{aligned} \tag{18}$$

Now, since D and n are at right angles,

$$D_1 n_1 + D_2 n_2 + D_3 n_3 = 0. \tag{19}$$

Hence,

$$\begin{aligned} 0 = {} & [(\beta_{22} - v^2/V^2)(\beta_{33} - v^2/V^2) - \beta_{23}^2] n_1^2 \\ & + [(\beta_{11} - v^2/V^2)(\beta_{33} - v^2/V^2) - \beta_{13}^2] n_2^2 \\ & + [(\beta_{11} - v^2/V^2)(\beta_{22} - v^2/V^2) - \beta_{12}^2] n_3^2. \end{aligned} \tag{20}$$

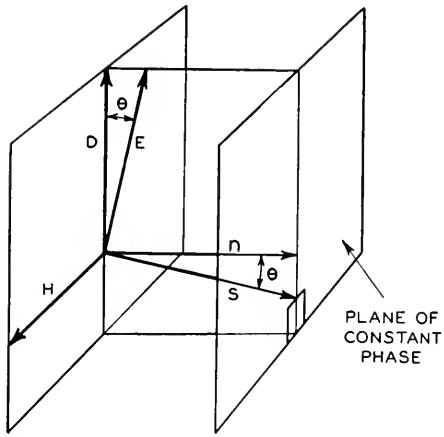

Fig. 1—Position of electric, magnetic and normal vectors for an electromagnetic plane wave in a crystal.

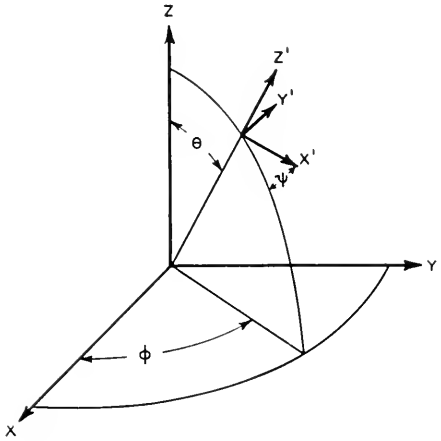

Fig. 2—Rotated axes and angles for relating them to unrotated axes.

By choosing the original x, y, z axes so that $\beta_{12} = \beta_{13} = \beta_{23} = 0$ and using the values $\beta_{11} = \beta_1$, $\beta_{22} = \beta_2$, $\beta_{33} = \beta_3$ this gives the equation

$$\frac{n_1^2}{\beta_1 - \frac{v^2}{V^2}} + \frac{n_2^2}{\beta_2 - \frac{v^2}{V^2}} + \frac{n_3^2}{\beta_3 - \frac{v^2}{V^2}} = 0. \tag{21}$$

For transmission along the X axis $n_1 = 1$, $n_2 = n_3 = 0$ and the two velocities are given by

$$v^2 = \beta_2 V^2 = b^2, \qquad v^2 = \beta_3 V^2 = c^2. \tag{22}$$

Similarly the third velocity $v^2 = \beta_1 V^2 = a^2$ can also be used and equation (21) reduces to

$$\frac{n_1^2}{a^2 - v^2} + \frac{n_2^2}{b^2 - v^2} + \frac{n_3^2}{c^2 - v^2} = 0. \tag{23}$$

This is a quadratic equation for the velocities v in terms of the principal velocities a, b and c which are usually taken so that $a > b > c$.

Solving for the velocities, we obtain the quadratic equation

$$v^4 - v^2[n_1^2(b^2 + c^2) + n_2^2(a^2 + c^2) + n_3^2(a^2 + b^2)] + n_1^2 b^2 c^2 + n_2^2 a^2 c^2 + n_3^2 a^2 b^2 = 0. \tag{24}$$

Letting $L = n_1^2(b^2 - c^2)$, $M = n_2^2(c^2 - a^2)$, $N = n_3^2(a^2 - b^2)$ the solutions for the velocities become

$$2v^2 = n_1^2(b^2 + c^2) + n_2^2(c^2 + a^2) + n_3^2(a^2 + b^2) \pm \sqrt{L^2 + M^2 + N^2 - 2LM - 2LN - 2MN}. \tag{25}$$

This equation can be put into a simpler form if we change to the coordinate system shown by Fig. 2. Here the rotated system is related to the original system by three angles θ, φ, ψ. θ is the angle between the Z' axis and the Z axis, φ is the angle the plane containing Z and Z' makes with the X axis while ψ represents a rotation of the primed coordinate systems about the Z' axis. The direction cosines for the primed system with respect to the normal system are designated by the matrix

	X	Y	Z
X'	ℓ_1	m_1	n_1
Y'	ℓ_2	m_2	n_2
Z'	ℓ_3	m_3	n_3

(26)

where, in terms of θ, φ and ψ, these direction cosines are,

$$\ell_1 = \cos\theta\cos\varphi\cos\psi - \sin\varphi\sin\psi,$$

$$m_1 = \cos\theta\sin\varphi\cos\psi + \cos\varphi\sin\psi, \qquad n_1 = -\sin\theta\cos\psi$$

$$\ell_2 = -\cos\theta\cos\varphi\sin\psi - \sin\varphi\cos\psi,$$

$$m_2 = \cos\varphi\cos\psi - \sin\varphi\sin\psi\cos\theta, \qquad n_2 = \sin\theta\sin\psi$$

$$\ell_3 = \cos\varphi\sin\theta, \qquad m_3 = \sin\varphi\sin\theta, \qquad n_3 = \cos\theta. \tag{27}$$

If we take Z' as the direction of the wave normal, then in equation (25)

$$n_1 = \ell_3, \qquad n_2 = m_3, \qquad n_3 = n_3$$

and the equation for the velocities becomes

$$2v^2 = a^2(\sin^2\varphi\sin^2\theta + \cos^2\theta) + b^2(\cos^2\varphi\sin^2\theta + \cos^2\theta) + c^2\sin^2\theta \pm \sqrt{(a^2 - b^2)^2(\cos^2\theta\cos^2\varphi + \sin^2\varphi)^2 + 2(a^2 - b^2)(c^2 - b^2)\sin^2\theta(\cos^2\theta\cos^2\varphi - \sin^2\varphi) + (c^2 - b^2)^2\sin^4\theta} \quad (28)$$

A very elegant construction for the wave-velocities and the directions of vibration is the Fresnel index ellipsoid. Consider the ellipsoid

$$a^2x^2 + b^2y^2 + c^2z^2 = 1 \quad (29)$$

Then Fresnel[3] showed that, for any diametral plane perpendicular to the wave normal, the two principal axes of the ellipse were the directions of the two permitted vibrations, while the wave velocities were the reciprocals of the principal semi-axes.

We wish to show now that the maximum and minimum values of the impermeability constants in a plane perpendicular to the direction of the wave normal determine the directions of vibration and the values of the two velocities. To show this we make use of the fact that β_{ij} is a second rank tensor and transforms according to the tensor transformation formula

$$\beta'_{ij} = \frac{\partial x'_i}{\partial x_k}\frac{\partial x'_j}{\partial x_\ell}\beta_{k\ell} \quad (30)$$

where the partial derivatives are the direction cosines

$$\frac{\partial x'_1}{\partial x_1} = \ell_1, \quad \frac{\partial x'_1}{\partial x_2} = m_1, \quad \frac{\partial x'_1}{\partial x_3} = n_1$$

$$\frac{\partial x'_2}{\partial x_1} = \ell_2, \quad \frac{\partial x'_2}{\partial x_2} = m_2, \quad \frac{\partial x'_2}{\partial x_3} = n_2$$

$$\frac{\partial x'_3}{\partial x_1} = \ell_3, \quad \frac{\partial x'_3}{\partial x_2} = m_3, \quad \frac{\partial x'_3}{\partial x_3} = n_3.$$

Expanding equation (30) the six transformation equations become

$$\beta'_{11} = \ell_1^2\beta_{11} + 2\ell_1m_1\beta_{12} + 2\ell_1n_1\beta_{13} + m_1^2\beta_{22} + 2m_1n_1\beta_{23} + n_1^2\beta_{33}$$

$$\beta'_{12} = \ell_1\ell_2\beta_{11} + (\ell_1m_2 + m_1\ell_2)\beta_{12} + (\ell_1n_2 + n_1\ell_2)\beta_{13} + m_1m_2\beta_{22} + (m_1n_2 + n_1m_2)\beta_{23} + n_1n_2\beta_{33}$$

$$\beta'_{13} = \ell_1\ell_3\beta_{11} + (\ell_1m_3 + m_1\ell_3)\beta_{12} + (\ell_1n_3 + n_1\ell_3)\beta_{13} + m_1m_3\beta_{22} + (n_1m_3 + m_1n_3)\beta_{23} + n_1n_3\beta_{33} \quad (31)$$

[3] See for example "Photoelasticity," Coker and Filon, Cambridge University Press, pages 17 and 18.

$$\beta'_{22} = \ell_2^2\beta_{11} + 2\ell_2 m_2\beta_{12} + 2\ell_2 n_2\beta_{13} + m_2^2\beta_{22} + 2m_2 n_2\beta_{23} + n_2^2\beta_{33}$$

$$\beta'_{23} = \ell_2\ell_3\beta_{11} + (\ell_2 m_3 + m_2\ell_3)\beta_{12} + (\ell_2 n_3 + n_2\ell_3)\beta_{13} + m_2 m_3\beta_{22} + (m_2 n_3 + n_2 m_3)\beta_{23} + n_2 n_3\beta_{33}$$

$$\beta'_{33} = \ell_3^2\beta_{11} + 2\ell_3 m_3\beta_{12} + 2\ell_3 n_3\beta_{13} + m_3^2\beta_{22} + 2m_3 n_3\beta_{23} + n_3^2\beta_{33}.$$

Now, if the axes refer to the axes of a Fresnel ellipsoid, $\beta_{12} = \beta_{13} = \beta_{23} = 0$ and one of the impermeability constants for any direction, say β'_{33}, can be expressed in the form

$$\beta'_{33} = \ell_3^2\beta_1 + m_3^2\beta_2 + n_3^2\beta_3 \tag{32}$$

If r, which lies along Z' of Fig. 2, is the radius vector of the Fresnel ellipsoid, then the direction cosines ℓ_3, m_3 and n_3 are

$$\ell_3 = \frac{x}{r}, \qquad m_3 = \frac{y}{r}, \qquad n_3 = \frac{z}{r}.$$

From equation (24) $\beta_1 = a^2/V^2$, $\beta_2 = b^2/V^2$, $\beta_3 = c^2/V^2$ and equation (32) becomes

$$r^2V^2\beta'_{33} = a^2x^2 + b^2y^2 + c^2z^2 = 1.$$

Hence the square of the radius vector of the Fresnel ellipsoid is $1/V^2\beta'_{33}$ and the radius vector of the impermeability ellipsoid agrees with that of the Fresnel ellipsoid. Hence, the directions of vibration can be determined from the principal axes of the impermeability ellipsoid for any diametral plane.

When light transmission occurs along Z', the direction for maximum and minimum impermeability can be obtained by evaluating β'_{11} and determining the angle ψ for which it has an extreme value. Inserting the direction cosines ℓ_1, m_1 and n_1 from equation (27), we find

$$\begin{aligned}\beta'_{11} = {} & \beta_1\left[\cos^2\theta\cos^2\varphi\cos^2\psi - \frac{\sin 2\varphi\sin 2\psi\cos\theta}{2} + \sin^2\varphi\sin^2\psi\right] \\ & + \beta_2\left[\cos^2\theta\sin^2\varphi\cos^2\psi + \frac{\sin 2\varphi\sin 2\psi\cos\theta}{2} + \cos^2\varphi\sin^2\psi\right] \\ & + \beta_3\sin^2\theta\cos^2\psi.\end{aligned} \tag{33}$$

Differentiating with respect to ψ and setting the resultant derivative equal to zero, the value of ψ that will satisfy the equation is given by

$$\begin{aligned}\tan 2\psi &= \frac{(\beta_2 - \beta_1)\sin 2\varphi\cos\theta}{(\beta_1 - \beta_2)(\cos^2\theta\cos^2\varphi - \sin^2\varphi) + (\beta_3 - \beta_2)\sin^2\theta} \\ &= \frac{(b^2 - a^2)\sin 2\varphi\cos\theta}{(a^2 - b^2)(\cos^2\theta\cos^2\varphi - \sin^2\varphi) + (c^2 - b^2)\sin^2\theta}.\end{aligned} \tag{34}$$

For a given value on the right-hand side there are two values of ψ, 90° apart, that will satisfy the equation and hence we have two directions of vibration at right angles to each other. Inserting (34) in (33) the values of β'_{11} and β''_{11} for these two directions are

$$2\beta'_{11} = \beta_1(\sin^2\varphi \sin^2\theta + \cos^2\theta) + \beta_2(\cos^2\varphi \sin^2\theta + \cos^2\theta) + \beta_3 \sin^2\theta \pm \sqrt{(\beta_1 - \beta_2)^2 (\cos^2\theta \cos^2\varphi + \sin^2\varphi)^2 + 2(\beta_1 - \beta_2)(\beta_3 - \beta_2) \cdot \sin^2\theta (\cos^2\theta \cos^2\varphi - \sin^2\varphi) + (\beta_3 - \beta_2)^2 \sin^4\theta}.$$

Since β_1 corresponds to a^2, etc., this equation agrees with the two velocities given in equation (28) and shows that the directions of vibration correspond with the maximum and minimum values of β'_{11}.

It can also be shown that the two directions of electric displacement coincide with the two values of ψ given by equation (34). Transforming the electrical displacements to the X', Y', Z' set of axes we have

$$\begin{aligned} D'_1 &= \frac{\partial x'_1}{\partial x_1} D_1 + \frac{\partial x'_1}{\partial x_2} D_2 + \frac{\partial x'_1}{\partial x_3} D_3 = \ell_1 D_1 + m_1 D_2 + n_1 D_3 \\ D'_2 &= \frac{\partial x'_2}{\partial x_1} D_1 + \frac{\partial x'_2}{\partial x_2} D_2 + \frac{\partial x'_2}{\partial x_3} D_3 = \ell_2 D_1 + m_2 D_2 + n_2 D_3 \\ D'_3 &= \frac{\partial x'_3}{\partial x_1} D_1 + \frac{\partial x'_3}{\partial x_2} D_2 + \frac{\partial x'_3}{\partial x_3} D_3 = \ell_3 D_1 + m_3 D_2 + n_3 D_3. \end{aligned} \tag{35}$$

Hence, inserting the values of D_1, D_2, D_3 from equation (18), we find

$$\begin{aligned} D'_1 &= \ell_1\ell_3(\beta_2 - \beta'_{11})(\beta_3 - \beta'_{11}) + m_1m_3(\beta_1 - \beta'_{11})(\beta_3 - \beta'_{11}) + n_1n_3(\beta_1 - \beta'_{11})(\beta_2 - \beta'_{11}) \\ D'_2 &= \ell_2\ell_3(\beta_2 - \beta'_{11})(\beta_3 - \beta'_{11}) + m_2m_3(\beta_1 - \beta'_{11})(\beta_3 - \beta'_{11}) + n_2n_3(\beta_1 - \beta'_{11})(\beta_2 - \beta'_{11}) \\ D'_3 &= \ell_3^2(\beta_2 - \beta'_{11})(\beta_3 - \beta'_{11}) + m_3^2(\beta_1 - \beta'_{11})(\beta_3 - \beta'_{11}) + n_3^2(\beta_1 - \beta'_{11})(\beta_2 - \beta'_{11}). \end{aligned} \tag{36}$$

From equation (20) with $\beta_{12} = \beta_{13} = \beta_{23} = 0$, it is evident that the D_3 component vanishes and hence the two values of electric displacement lie in a plane perpendicular to Z'. By inserting the values of β'_{11} and the value of ψ found from equation (34) we find that $D_2 = 0$ and hence the electric displacement lies along the directions of the greatest value of β'_{11}. Similarly, from the second value of β'_{11}, D_1 vanishes and hence the second wave is perpendicular to the first and in the direction of the smallest value of β'_{11}.

III. Location of Optic Axes in a Crystal

When the expression in the radical of equation (28) vanishes the two velocities are equal and an optic axis exists. Since the expression inside the radical can be written

$$[(a^2 - b^2)(\cos^2\theta\cos^2\varphi + \sin^2\varphi) - (b^2 - c^2)\sin^2\theta]^2 - 4(a^2 - b^2)(c^2 - b^2)\sin^2\theta\sin^2\varphi = 0 \tag{37}$$

then, since the square is always positive and since $(a^2 - b^2) > 0$ and $(b^2 - c^2) > 0$, the equation can vanish only if $\varphi = 0$. But $\varphi = 0$ indicates that the two optic axes always lie in a plane perpendicular to the intermediate velocity b. With $\varphi = 0$ then the square vanishes when

$$\tan^2\theta = \frac{(a^2 - b^2)}{(b^2 - c^2)} \quad \text{or} \quad \tan\theta = \pm\sqrt{\frac{a^2 - b^2}{b^2 - c^2}}. \tag{38}$$

If $(a^2 - b^2) < (b^2 - c^2)$ the value of the $\tan\theta$ is less than unity and the crystal is called a positive crystal. For this case the two axes approach more closely the Z axis having the velocity c than they do the X axis. If $(a^2 - b^2) > (b^2 - c^2)$ the crystal is negative.

If $a = b$ or $b = c$ the crystal has a single optic axis and is respectively a positive or negative uniaxial crystal. For the first case the two velocities are given by

$$v_1 = a = b, \qquad v_2 = \sqrt{a^2\cos^2\theta + c^2\sin^2\theta}. \tag{39}$$

The first velocity is that of the ordinary ray while that of the second is that of the extraordinary ray. Since $a > c$, the ordinary ray will have a velocity greater than the extraordinary ray except along the optic axis where they are equal. Since $c < a$, the maximum axis for any ellipse, formed by intersecting the Fresnel ellipsoid at an angle to the optic axis, will lie in the plane formed by the normal and the c axis and hence the direction of polarization of the extraordinary ray will lie in the c, n plane. The polarization of the ordinary ray will be perpendicular to this plane.

If $b = c$ the a axis is the optic axis and the velocities of the two rays are again

$$v_1 = c \text{ and } v_2^2 = a^2(1\text{-}\sin^2\theta\cos^2\varphi) + c^2(\sin^2\theta\cos^2\varphi) \tag{40}$$

Hence, when $\theta = 90°$, $\varphi = 0°$, the two velocities are equal and a is the optic axis. In this case the velocity of the extraordinary ray is greater than that of the ordinary ray except along the a axis, and the crystal is a negative uniaxial crystal. The polarization of the extraordinary ray lies again in the

plane of the normal and the optic axis while the ordinary ray is perpendicular to it.

IV. Derivation of the Electro-optic and Photoelastic Effects

In a previous paper[4] and in the book "Piezoelectric Crystals and Their Application to Ultrasonics", D. Van Nostrand, 1950, it was shown that the electro-optic and photoelastic effects can be expressed as third derivatives of one of the thermodynamic potentials. Probably the most fundamental way of developing these properties is to express them in terms of the strains, electric displacements and the entropy. For viscoelastic substances it has been shown that the photoelastic effects are directly related to the strains. In terms of the electric displacements, the electro-optic constants do not vary much with temperature whereas, if they are expressed in terms of the fields, the constants of a ferroelectric type of crystal such as KDP increase many fold near the Curie temperature. The entropy is chosen as the fundamental heat variable, since most measurements are carried out so rapidly that the entropy does not vary.

The thermodynamic potential which has the strains, electric displacements and entropy as the independent variables is the internal energy U, given by

$$dU = T_{ij}\, dS_{ij} + E_m \frac{dD_m}{4\pi} + \Theta\, d\sigma \tag{41}$$

where S_{ij} are the strains, T_{ij} the stresses, E_m the fields, D_m the electric displacements, Θ the temperature and σ the entropy. In this equation the strains S_{ij} are defined in the tensor form

$$S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{42}$$

where the u's are the displacements along the three axis. In the case of a shearing strain occurring when $i \neq j$, the strain is only half that usually used in engineering practice. In order to avoid writing the factor $1/4\pi$, we use the variable $\delta_m = D_m/4\pi$. Then, from (41),

$$T_{ij} = \frac{\partial U}{\partial S_{ij}}, \qquad E_m = \frac{\partial U}{\partial \delta_n}, \qquad \theta = \frac{\partial U}{\partial \sigma}. \tag{43}$$

Since, for most conditions of interest, adiabatic conditions prevail, we can set $d\sigma$ equal to zero and can develop the dependent variables, the fields and

[4] "First and Second Order Equations for Piezoelectric Crystals Expressed in Tensor Form," W. P. Mason, *B.S.T.J.*, Vol. 26, pp. 80–138, Jan., 1947.

the stresses in terms of the independent variables, the strains and the electric displacements. Up to the second derivatives, these are

$$\begin{aligned} E_m &= \frac{\partial E_m}{\partial S_{ij}} S_{ij} + \frac{\partial E_m}{\partial \delta_n} \delta_n \\ &\quad + \frac{1}{2!}\left[\frac{\partial^2 E_m}{\partial S_{ij}\,\partial S_{qr}} S_{ij} S_{qr} + \frac{2\partial^2 E_m}{\partial S_{ij}\,\partial \delta_n} S_{ij}\delta_n + \frac{\partial^2 E_m}{\partial \delta_n\,\partial \delta_o}\delta_n \delta_o\right] + \cdots \\ T_{k\ell} &= \frac{\partial T_{k\ell}}{\partial S_{ij}} S_{ij} + \frac{\partial T_{k\ell}}{\partial \delta_n}\delta_n \\ &\quad + \frac{1}{2!}\left[\frac{\partial^2 T_{k\ell}}{\partial S_{ij}\,\partial S_{qr}} S_{ij} S_{qr} + \frac{2\partial^2 T_{k\ell}}{\partial S_{ij}\,\partial \delta_n} S_{ij}\delta_n + \frac{\partial^2 T_{k\ell}}{\partial \varepsilon_n\,\partial \delta_o}\delta_n \delta_o\right] + \cdots . \end{aligned} \tag{44}$$

For the electro-optic and photoelastic cases, the two tensors of interest are

$$\begin{aligned} \frac{\partial^2 T_{k\ell}}{\partial \delta_n\,\partial \delta_o} &= \frac{\partial^3 U}{\partial S_{k\ell}\,\partial \delta_n\,\partial \delta_o} = \frac{\partial^2 E_n}{\partial S_{k\ell}\,\partial \delta_o} = 4\pi m_{k\ell no} \\ \frac{\partial^2 E_m}{\partial \delta_n\,\partial \delta_o} &= \frac{\partial^3 U}{\partial \delta_m\,\partial \delta_n\,\partial \delta_o} = (4\pi) r_{mno} . \end{aligned} \tag{45}$$

For the first partial derivatives, we have the values

$$\begin{aligned} \frac{\partial T_{k\ell}}{\partial S_{ij}} &= c^D_{ijk\ell} ; \qquad \frac{\partial T_{k\ell}}{\partial \delta_n} = \frac{\partial^2 U}{\partial S_{k\ell}\,\partial \delta_n} = \frac{\partial E_n}{\partial S_{k\ell}} = -h_{nk\ell} \\ &\qquad \frac{\partial E_m}{\partial \delta_n} = 4\pi \beta^S_{mn} \end{aligned} \tag{46}$$

where $c^D_{ijk\ell}$ are the elastic stiffnesses measured at constant electric displacement, $h_{nk\ell}$ are the piezoelectric constants that relate the open circuit voltages to the strains, and β^S_{mn} are the impermeability constants measured for constant strain.

With these substitutions and neglecting the other second partial derivatives, we have, from (44),

$$\begin{aligned} E_m &= -h_{mij} S_{ij} + D_n\left[\beta^S_{mn} + m_{ijmn} S_{ij} + \frac{r^S_{mno}}{2} D_o\right] + \cdots \\ T_{k\ell} &= c^D_{ijk\ell} S_{ij} + D_o\left[-\frac{h_{ok\ell}}{4\pi} + \frac{m_{k\ell on} D_n}{2}\right]. \end{aligned} \tag{47}$$

This equation shows that there is a relation between the change in the impermeability constant due to stress in the first equation, and the electrostrictive constant in the second equation through the tensor m_{ijmn}. These

effects, however, have to be measured at the same frequency before equality exists.

To obtain the changes in the optical properties caused by the strain and the electric displacement we have to determine the fields and displacements occurring at the high frequencies of optics. Even for piezoelectric vibrations occurring at as high frequencies as they can be driven by the piezoelectric effect, these frequencies are small compared to the optic frequencies f and can be considered to be static displacements or strains. Hence, writing

$$E_m = E_m^0 + E_m e^{j\omega t}, \qquad D_n = D_n^0 + D_n e^{j\omega t},$$
$$D_o = D_o^0 + D_o e^{j\omega t}, \qquad S_{ij} = S_{ij}^0$$

where $\omega = 2\pi f$, the first of equation (47) can be written in the form

$$E_m^0 = -h_{mij} S_{ij} + D_n^0 \left[\beta_{mn}^S + m_{ijmn} S_{ij} + \frac{r_{mno}}{2} D_o^0 \right]$$
$$E_m e^{j\omega t} = D_n e^{j\omega t} \left[\beta_{mn}^S + m_{ijmn} S_{ij} + \frac{r_{mno}^S}{2} D_o^0 \right] + \frac{r_{mno}}{2} D_n^0 D_o e^{j\omega t}. \tag{48}$$

If we develop one of the fields, say E_1, this can be written in the form

$$\begin{aligned} E_1 e^{j\omega t} = {} & [\beta_{11} + m_{ij11} S_{ij} + r_{111} D_1^0 + r_{112} D_2^0 + r_{113} D_3^0] D_1 e^{j\omega t} \\ & + [\beta_{12} + M_{ij12} S_{ij} + r_{121} D_1^0 + r_{122} D_2^0 + r_{123} D_3^0] D_2 e^{j\omega t} \\ & + [\beta_{13} + M_{ij13} S_{ij} + r_{131} D_1^0 + r_{132} D_2^0 + r_{133} D_3^0] D_3 e^{i\omega t} \end{aligned} \tag{49}$$

where the first number of r refers to the field, the second to the optical value of D and the third to the static value of D. Hence, for the general case,

$$E_m e^{j\omega t} = D_n e^{j\omega t} [\beta_{mn} + m_{ijmn} S_{ij} + r_{mno} D_o^0]. \tag{50}$$

From the definition of the two tensors m_{ijno} and r_{mno} given by equation (45), we can show that there are relations between the various components of the tensors. For the first tensor m_{ijno}, since $S_{ij} = S_{ji}$ is a symmetrical tensor, then

$$m_{ijno} = m_{jino} \tag{51}$$

From the definition of the tensor m_{ijno} in the form

$$4\pi m_{ijno} = \frac{\partial}{\partial S_{ij}} \left(\frac{\partial^2 U}{\partial \delta_n \, \partial \delta_o} \right) \tag{45}$$

it is obvious that we can interchange the order of δ_n and δ_o so that

$$m_{ijno} = m_{ijon}$$

Since ij and no are reversible, it has been customary to abbreviate the tensor by writing one number in place of the two in the following form:

$$11 = 1; 22 = 2; 33 = 3; 12 = 21 = 6; 13 = 31 = 5; 23 = 32 = 4 \quad (52)$$

Since the reduced tensor is associated with the engineering strains, it is necessary to investigate the numerical relationships between the four index symbols and the two index symbols. From equation (48), when $m \neq n$, the change in the impermeability constant β_{mn} is given by

$$m_{ijmn} S_{ij} + m_{jimn} S_{ji} = m_{rs} S_r \quad (53)$$

Since $S_r = 2S_{ij} = 2S_{ji}$ we have the relation that

$$m_{ijmn} = m_{rs} (i, j, m, n = 1 \text{ to } 3, r, s, = 1 \text{ to } 6) \quad (54)$$

In equation (45) we cannot in general interchange the order of ij and no since U does not contain product terms of strains and electric displacements and hence in general

$$m_{rs} \neq m_{sr} . \quad (55)$$

Hence in the most general case there are 36 photoelastic constants. Crystal symmetrics cut down the number of constants as shown in a later section.

The tensor r_{mno} defined in equation (45) as

$$(4\pi)^2 r_{mno} = \frac{\partial^3 U}{\partial \delta_m \partial \delta_n \partial \delta_o} \quad (56)$$

shows that we can interchange the order of m and n since U contains product terms of δ_m and δ_n. Hence

$$r_{mno} = r_{nmo} \quad (57)$$

and this is usually replaced by the two index symbols

$$r_{qo} = r_{mno} (m, n, o = 1 \text{ to } 3; q = 1 \text{ to } 6).$$

The so called "true" electro-optic constants are measured at constant strain and for this case the modifications in the impermeability constants are given by the equation

$$E_m = D_n [\beta^S_{mn} + r^S_{mno} D_o] . \quad (58)$$

Since m and n are interchangeable, the third rank tensor is usually replaced by the two index symbols

$$r^S_{mno} = r_{qo} (m, n, o = 1 \text{ to } 3; q = 1 \text{ to } 6). \quad (59)$$

As discussed in the next sections, these constants can be determined by applying an electric field of a frequency high enough so that the principal resonances and their harmonics cannot be excited by the applied field, and measuring the resulting birefringence along definite directions in the crystal. On the other hand if we apply a static field to the crystal, an additional effect occurs because the crystal is strained by the piezoelectric effect and this causes a photoelastic effect in addition to the "true" electro-optic effect. A

better designation for these effects is the electro-optic effect at constant strain and stress.

This latter effect can be calculated from equation (47) by setting the stresses $T_{k\ell}$ equal to zero and eliminating the S_{ij} strains. After neglecting second order corrections,

$$E_m = D_n e^{j\omega t}\left[\beta^S_{mn} + \left(r^S_{mno} + \frac{m_{ijmn}\, h_{ok\ell}}{4\pi c^D_{ijk\ell}}\right) D^0_o\right]. \tag{60}$$

Since $h_{ok\ell}/c^D_{ijk\ell} = g_{oij}$, the other piezoelectric constant relating the open circuit voltage to the stress, the electro-optic effect at constant stress can be written in the form

$$r^T_{mno} = r^S_{mno} + \frac{m_{ijmn}\, g_{oij}}{4\pi}. \tag{61}$$

In terms of the two index symbols

$$r^T_{qo} = r^S_{qo} + \frac{m_{pq}\, g_{op}}{4\pi} \tag{62}$$

since it has been shown[4] that $g_{oij} = g_{op}/2$ when $i \neq j$, and the tensor in (61) has ij as common symbols which involves the summations of two terms.

The electro-optic effect is usually measured in terms of an applied field. The change in the impermeability constant β^S_{mn} for this case can be determined from the first equations (47), setting $T_{k\ell}$ equal to zero and neglecting second order terms. Multiplying through by the tensor K^T_{op} of the dielectric constants

$$D^0_p = E^0_0\, K^T_{op} \tag{63}$$

since the product $K^T_{op}\beta^T_{op} = 1$. Introducing this equation into (58) we have

$$E_m = D_n[\beta^S_{mn} + r^S_{mnp} K^T_{op} E^0_o] = D_n[\beta^S_{mn} + z^S_{mno} E^0_o]. \tag{64}$$

where the new tensor z_{mno} is equal to

$$z^S_{mno} = r^S_{mnp} K^T_{op}. \tag{65}$$

In terms of the two index symbols

$$z^S_{qo} = r^S_{qp} K^T_{op}. \tag{66}$$

in which the repeated index indicates a summation. The difference between the electro-optical constant at constant stress expressed in terms of the field and the electro-optical constant at constant strain is

$$z^T_{mno} = z^S_{mno} + \frac{m_{ijmn}\, g_{oij}}{4\pi} K^T_{op} = z^S_{mno} + m_{ijmn}\, d_{pij} \tag{67}$$

since the piezoelectric constants d_{pij} are related to the g constants by the equation

$$d_{pij} = \frac{g_{oij} K^T_{op}}{4\pi}. \tag{68}$$

In terms of two index symbols

$$z_{qo}^{T} = z_{qo}^{S} + m_{pq}d_{op}(p, q = 1 \text{ to } 6; o = 1 \text{ to } 3) \tag{69}$$

where a repeated index means a summation with respect to this index.

Finally the photoelastic effect is sometimes expressed in terms of the stresses rather than the strains. As can be seen from equation (47), the new set of constants is

$$\pi_{pq} = m_{pr}s_{rq}^{D} \tag{70}$$

where the s_{rq}^{D} are the elastic compliances measured at constant electric displacement.

V. Birefringence Along Any Direction In the Crystal and Determination of the Electro-optic and Photoelastic Constants

If we take axes along the Fresnel ellipsoid when no stress or field is applied to the crystal, the result of the electro-optic and photoelastic effects is to change the impermeability constants by the values

$$\begin{aligned} &\beta_{11} = \beta_1 + \Delta_1 ; \quad \beta_{22} = \beta_2 + \Delta_2 ; \quad \beta_{33} = \beta_3 + \Delta_3 \\ &\beta_{23} = \Delta_4 ; \quad \beta_{13} = \Delta_5 ; \quad \beta_{12} = \Delta_6 \end{aligned} \tag{71}$$

where

$$\begin{aligned} \Delta_1 &= z_{11}E_1 + z_{12}E_2 + z_{13}E_3 + m_{11}S_1 + m_{12}S_2 + m_{13}S_3 + m_{14}S_4 \\ &\quad + m_{15}S_5 + m_{16}S_6 \\ \Delta_2 &= z_{21}E_1 + z_{22}E_2 + z_{23}E_3 + m_{21}S_1 + m_{22}S_2 + m_{23}S_3 + m_{24}S_4 \\ &\quad + m_{25}S_5 + m_{26}S_6 \\ \Delta_3 &= z_{31}E_1 + z_{32}E_2 + z_{33}E_3 + m_{31}S_1 + m_{32}S_2 + m_{33}S_3 + m_{34}S_4 \\ &\quad + m_{35}S_5 + m_{36}S_6 \\ \Delta_4 &= z_{41}E_1 + z_{42}E_2 + z_{43}E_3 + m_{41}S_1 + m_{42}S_2 + m_{43}S_3 + m_{44}S_4 \\ &\quad + m_{45}S_5 + m_{46}S_6 \\ \Delta_5 &= z_{51}E_1 + z_{52}E_2 + z_{53}E_3 + m_{51}S_1 + m_{52}S_2 + m_{53}S_3 + m_{54}S_4 \\ &\quad + m_{55}S_5 + m_{56}S_6 \\ \Delta_6 &= z_{61}E_1 + z_{62}E_2 + z_{63}E_3 + m_{61}S_1 + m_{62}S_2 + m_{63}S_3 + m_{64}S_4 \\ &\quad + m_{65}S_5 + m_{66}S_6 . \end{aligned} \tag{72}$$

If we transmit light along the z' axis which, as shown by Fig. 2, makes an angle of θ degrees with the z axis in a plane making an angle φ with the xz plane, the birefringence can be calculated as follows: Keeping z' fixed and rotating the other two axes about z' by varying the angle ψ, one light vector

will occur when β'_{11} is a maximum and the other when β'_{11} is a minimum. Using the transformation equations (31) and the direction cosines of (27), we find that β'_{11} is given by the equations

$$\begin{aligned}\beta'_{11} = {} & \beta_{11}\left[\cos^2\theta\cos^2\varphi\cos^2\psi - \frac{\sin 2\varphi\sin 2\psi\cos\theta}{2} + \sin^2\varphi\sin^2\psi\right] \\ & + \beta_{12}[\sin 2\varphi\cos 2\psi - \sin^2\theta\sin 2\varphi\cos^2\psi + \cos\theta\sin 2\psi\cos 2\varphi] \\ & + \beta_{13}[-\sin 2\theta\cos\varphi\cos^2\psi + \sin\varphi\sin\theta\sin 2\psi] \\ & + \beta_{22}\left[\cos^2\theta\sin^2\varphi\cos^2\psi + \frac{\cos\theta\sin 2\varphi\sin 2\psi}{2} + \cos^2\varphi\sin^2\psi\right] \\ & + \beta_{23}[-\sin 2\theta\sin\varphi\cos^2\psi - \sin\theta\cos\varphi\sin 2\psi] + \beta_{33}\sin^2\theta\cos^2\psi\end{aligned} \tag{73}$$

Differentiating with respect to ψ and setting $\dfrac{\partial\beta'_{11}}{\partial\psi} = 0$, we find an expression for $\tan 2\psi$ in the form

$$\tan 2\psi = \frac{\begin{array}{l}-\beta_{11}\sin 2\varphi\cos\theta + 2\beta_{12}\cos\theta\cos 2\varphi \\ + 2\beta_{13}\sin\varphi\sin\theta + \beta_{22}\cos\theta\sin 2\varphi - 2\beta_{23}\sin\theta\cos\varphi\end{array}}{\begin{array}{r}\beta_{11}[\cos^2\theta\cos^2\varphi - \sin^2\varphi] + \beta_{12}[(1+\cos^2\theta)\sin 2\varphi] \\ - \beta_{13}\sin^2\theta\cos\varphi + \beta_{22}(\cos^2\theta\sin^2\varphi - \cos^2\varphi) \\ - \beta_{23}\sin 2\theta\sin\varphi + \beta_{33}\sin^2\theta\end{array}} \tag{74}$$

Inserting this value back in equation (73) we find that the two extreme values of β'_{11} are given by the equation

$$\begin{aligned}2\beta'_{11} = {} & 2\beta_{22} + (\beta_{11} - \beta_{22})(\cos^2\theta\cos^2\varphi + \sin^2\varphi) + (\beta_{33} - \beta_{22})\sin^2\theta \\ & - \beta_{12}\sin^2\theta\sin 2\varphi - \beta_{13}\sin 2\theta\cos\varphi - \beta_{23}\sin 2\theta\sin\varphi \\ & \pm \sqrt{\begin{array}{l}(\beta_{11} - \beta_{22})^2(\cos^2\theta\cos^2\varphi + \sin^2\varphi)^2 + 2(\beta_{11} - \beta_{22})(\beta_{33} - \beta_{22})\sin^2\theta\times \\ (\cos^2\theta\cos^2\varphi - \sin^2\varphi) + (\beta_{33} - \beta_{22})^2\sin^4\theta - 2(\beta_{11} - \beta_{22})\times \\ [\beta_{12}(\sin 2\varphi\sin^2\theta(\cos^2\theta\cos^2\varphi + \sin^2\varphi) + \beta_{13}\sin 2\theta\cos\varphi\times \\ (\cos^2\theta\cos^2\varphi + \sin^2\varphi) - \beta_{23}\sin 2\theta\sin\varphi(1 + \cos^2\varphi\sin^2\theta)] \\ + 2(\beta_{33} - \beta_{22})\sin^2\theta[\beta_{12}\sin 2\varphi(1 + \cos^2\theta) - \beta_{13}\sin 2\theta\cos\varphi \\ - \beta_{23}\sin 2\theta\sin\varphi] + (2\beta_{12})^2[\sin^4\theta\sin^2\varphi\cos^2\varphi + \cos^2\theta] \\ - 4\beta_{12}\beta_{13}\sin^2\theta\sin\varphi[\cos^2\theta\cos^2\varphi + \sin^2\varphi] - 4(\beta_{12}\beta_{23}) \\ [\sin 2\theta\cos\varphi(\sin^2\varphi\cos^2\theta + \cos^2\varphi)] + (2\beta_{13})^2\sin^2\theta\times \\ (\cos^2\theta\cos^2\varphi + \sin^2\varphi) - 4\beta_{13}\beta_{23}\sin 2\varphi\sin^4\theta \\ + (2\beta_{23})^2\sin^2\theta(\cos^2\theta\sin^2\varphi + \cos^2\varphi)\end{array}}\end{aligned} \tag{75}$$

The birefringence in any direction can be calculated from equation (75); since $\beta'_{11} = v_1^2/V^2$, it equals $1/\mu_1^2$ where μ_1 is the index of refraction corresponding to a light wave with its electric displacement in the β'_{11} direction. Similarly, for the second solution at right angle to the first,

$$\beta''_{11} = \frac{v_2^2}{V^2} = \frac{1}{\mu_2^2} \tag{76}$$

Hence if we designate the expression under the radical by K_2 and half the expression on the right outside the radical by K_1, we have

$$\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} = K_1; \qquad \frac{1}{\mu_1^2} - \frac{1}{\mu_2^2} = \sqrt{K_2}. \tag{77}$$

Since μ_1 and μ_2 are very nearly equal even in the most birefringent crystal, we have nearly

$$\mu_2 - \mu_1 = B = \frac{\mu^3}{2}\sqrt{K_2}. \tag{78}$$

For special directions in the crystal, the expression for K_2 simplifies very considerably. Along the x, y and z axes, the values are

$$\begin{aligned} X,\ (\varphi = 0°, \theta = 90°); \qquad & B_x = \frac{\mu^3}{2}\sqrt{(\beta_{33} - \beta_{22})^2 + (2\beta_{23})^2} \\ Y,\ (\varphi = 90°, \theta = 90°); \qquad & B_y = \frac{\mu^3}{2}\sqrt{(\beta_{11} - \beta_{33})^2 + (2\beta_{13})^2} \\ Z,\ (\varphi = 0°, \theta = 0°); \qquad & B_z = \frac{\mu^3}{2}\sqrt{(\beta_{11} - \beta_{22})^2 + (2\beta_{12})^2}. \end{aligned} \tag{79}$$

If any natural birefringence exists along these axes, $(2\beta_{23})^2$ will be very small compared to this and

$$\begin{aligned} B_x &= \frac{\mu^3}{2}(\beta_3 - \beta_2 + \Delta_3 - \Delta_2) = \frac{\mu^3}{2}\left(\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2} + \Delta_3 - \Delta_2\right) \\ B_y &= \frac{\mu^3}{2}(\beta_1 - \beta_3 + \Delta_1 - \Delta_3) = \frac{\mu^3}{2}\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_c^2} + \Delta_1 - \Delta_3\right) \\ B_z &= \frac{\mu^3}{2}(\beta_1 - \beta_2 + \Delta_1 - \Delta_2) = \frac{\mu^3}{2}\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_b^2} + \Delta_1 - \Delta_2\right). \end{aligned} \tag{80}$$

Hence, for this case, measurements along the three axes will tell the difference between the three effects Δ_1, Δ_2 and Δ_3. To get absolute values requires a direct measurement of the index of refraction along one of the axes and its change with fields or stresses. This is a considerably more difficult meas-

urement than a birefringence measurement and requires the use of an accurate interferometer.

If, however, the Z axis is an optic axis as it is in ADP, for example, and $\Delta_1 = \Delta_2 = 0$, a birefringence occurs due to the term β_{12}. As shown in the next section, the electro-optic constants for ADP (tetragonal $\bar{4}2m$) are z_{41} and z_{63}. z_{63} occurs in the expression for $\beta_{12} = \Delta_6$, as can be seen from equations (72), and hence the birefringence along the Z axis is

$$B_z = \frac{\mu_a^3}{2} x 2\beta_{12} = \mu_a^3 z_{63} E_3. \tag{81}$$

The constants z_{63} and z_{41} have been measured independently by W. L. Bond, Robert O'B. Carpenter, and Hans Jaffe. Probably the most accurate measurements, and the only one published, are those of Carpenter,[5] who finds that the indices of refraction and the z_{63} and z_{41} constants for ADP and KDP are in cgs units

	μ_a	μ_c	$r_{63} \times 10^7$	$r_{41} \times 10^7$
ADP	1.5254	1.4798	2.54 ± 0.05	6.25 ± 0.1
KDP	1.5100	1.4684	3.15 ± 0.07	2.58 ± 0.05

An even larger constant has been found for heavy hydrogen KDP by Zwicker and Scherrer.[6] They find at 20°C that $r_{63} = 6 \times 10^{-7}$. Using this constant, a half wave retardation for a $\lambda = 5461$ A° mercury line occurs for a voltage of 4000 volts.

For tetragonal crystals of these types the only photoeleastic constant for the z axis is m_{66}, and the birefringence for this case is given by

$$B_z = \mu_a^3 m_{66} S_6 \tag{82}$$

When a natural birefringence exists for the crystal, measurements of the other three effects Δ_4, Δ_5 and Δ_6 can be made by determining the birefringence along other directions than the Fresnel ellipsoid axes. In a direction of Z' lying in the XZ plane $\varphi = 0$, $\theta =$ variable and

$$B_{zz} = \frac{\mu^3}{2} \sqrt{\begin{aligned}&[(\beta_{11} - \beta_{22}) \cos^2\theta + (\beta_{33} - \beta_{22}) \sin^2\theta - \beta_{13} \sin^2\theta]^2 \\ &+ [2\beta_{12} \cos\theta + 2\beta_{23} \sin\theta]^2.\end{aligned}} \tag{83}$$

When a natural birefringence exists, this reduces to

$$B_{xy} \doteq \frac{\mu^3}{2} \left[\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_b^2} + \Delta_1 - \Delta_2 \right) \cos^2\theta + \left(\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2} + \Delta_3 - \Delta_2 \right) \sin^2\theta - \Delta_5 \sin 2\theta \right] \tag{84}$$

[5] "The Electro-optic Effect in Uniaxial Crystals of the Type XH_2PO_4," Robert O'B. Carpenter, *Jour. Opt. Soc. Am.*, in course of publication.

[6] Zwicker and Scherrer, *Helv. Phys. Acta.*, 17, 346 (1944).

and hence, by measuring at 45° between the two axes, one can evaluate the Δ_5 term.

Similarly, for the YZ plane, $\varphi = 90°$, $\theta =$ variable and

$$B_{yz} = \frac{\mu^3}{2}\sqrt{[-(\beta_{11} - \beta_{22}) + (\beta_{33} - \beta_{22})\sin^2\theta - \beta_{23}\sin 2\theta]^2 + [2\beta_{12}\cos\theta - 2\beta_{13}\sin\theta]^2}. \tag{85}$$

Hence, when a natural birefringence exists, we have

$$B_{yz} \doteq \frac{\mu^3}{2}\left[-\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_b^2} + \Delta_1 - \Delta_2\right) + \left(\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2} + \Delta_3 - \Delta_2\right)\sin^2\theta - \Delta_4\sin 2\theta\right]. \tag{86}$$

In the XY plane $\theta = 90°$, $\varphi =$ variable and

$$B_{xy} = \frac{\mu^3}{2}\sqrt{[(\beta_{11} - \beta_{12})\sin^2\varphi - (\beta_{33} - \beta_{22}) - \beta_{12}\sin 2\varphi]^2 + [2\beta_{13}\sin\varphi - \beta_{23}\cos\varphi]^2}. \tag{87}$$

Then, for natural birefringence,

$$B_{xy} \doteq \frac{\mu^3}{2}\left[\left(\frac{1}{\mu_a^2} - \frac{1}{\mu_b^2} + \Delta_1 - \Delta_2\right)\sin^2\varphi - \left(\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2} + \Delta_3 - \Delta_2\right) - \Delta_6\sin 2\varphi\right]. \tag{88}$$

Hence, with measurements at 45° between the axes and with suitably applied fields and strains, the three effects Δ_4, Δ_5 and Δ_6 can be measured. Since the axes of the test specimen are turned with respect to the X, Y and Z axes, suitable transformations of the effects Δ_1 to Δ_6 with respect to the new axes will have to be made. These can be done as shown in reference (4) by means of tensor transformation formulae.

Another method for measuring the constants in Δ_4, Δ_5, Δ_6 is to measure the amount they rotate the axes of the Fresnel ellipsoid. As an example consider the z_{41} constant of ADP. For example, if we look along the X axis and apply a field in the same direction, then, in equation (74), $\theta = 90°$, $\varphi = 0$ and

$$\tan 2\psi = \frac{-2\beta_{23}}{\beta_{33} - \beta_{22}} = \frac{-2z_{41}E_1}{\frac{1}{\mu_c^2} - \frac{1}{\mu_b^2}} = \frac{-2\mu_b^2\mu_c^2 z_{41}E_1}{(\mu_b + \mu_c)(\mu_b - \mu_c)}. \tag{89}$$

According to Carpenter, the z_{41} electro-optic constant of ADP is 6.25×10^{-7} in cgs units. $\mu_a = \mu_b = 1.5254$; $\mu_c = 1.4798$; hence the angle of rotation for a field of 30,000 volts per centimeter = 100 stat volts cm is

$$\psi = -2.25 \times 10^{-3} \text{ radians} = 7.7 \text{ minutes of arc}. \tag{90}$$

VI. Electro-optic and Photoelastic Tensors for Various Crystal Classes

Since $r_{mno} = r_{nmo}$ and $z_{mno} = z_{nmo}$ are third rank tensors similar to the h_{mij} piezoelectric tensor, they will have the same components for the various crystal classes. For the twenty crystal classes that show the electro-optic effect these tensors are given below. They are given with the crystal system they belong to, and the symmetry is designated by the Hermann-Mauguin symbol. The last number of the subscript of z designates the direction of the applied static field.

(91)

Triclinic; 1	z_{11}	z_{21}	z_{31}	z_{41}	z_{51}	z_{61}
	z_{12}	z_{22}	z_{32}	z_{42}	z_{52}	z_{62}
	z_{13}	z_{23}	z_{33}	z_{43}	z_{53}	z_{63}
Monoclinic; 2	0	0	0	z_{41}	0	z_{61}
	z_{12}	z_{22}	z_{32}	0	z_{52}	0
	0	0	0	z_{43}	0	z_{63}
Monoclinic; $\bar{2} = m$	z_{11}	z_{21}	z_{31}	0	z_{51}	0
	0	0	0	z_{42}	0	z_{62}
	z_{13}	z_{23}	z_{33}	0	z_{53}	0
Orthorhombic; 222	0	0	0	z_{41}	0	0
	0	0	0	0	z_{52}	0
	0	0	0	0	0	z_{63}
Orthorhomic; 2mm	0	0	0	0	z_{51}	0
	0	0	0	z_{42}	0	0
	z_{13}	z_{23}	z_{33}	0	0	0
Tetragonal; 4	0	0	0	z_{41}	z_{51}	0
	0	0	0	$-z_{51}$	z_{41}	0
	z_{13}	$-z_{13}$	0	0	0	z_{63}

Tetragonal; 4

$$\begin{vmatrix} 0 & 0 & 0 & z_{41} & z_{51} & 0 \\ 0 & 0 & 0 & z_{51} & -z_{41} & 0 \\ z_{13} & z_{13} & z_{33} & 0 & 0 & 0 \end{vmatrix}$$

Tetragonal; $\bar{4}$2m

$$\begin{vmatrix} 0 & 0 & 0 & z_{41} & 0 & 0 \\ 0 & 0 & 0 & 0 & z_{41} & 0 \\ 0 & 0 & 0 & 0 & 0 & z_{63} \end{vmatrix}$$

Tetragonal; 422

$$\begin{vmatrix} 0 & 0 & 0 & z_{41} & 0 & 0 \\ 0 & 0 & 0 & 0 & -z_{41} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Tetragonal; 4mm

$$\begin{vmatrix} 0 & 0 & 0 & 0 & z_{51} & 0 \\ 0 & 0 & 0 & z_{51} & 0 & 0 \\ z_{13} & z_{13} & z_{33} & 0 & 0 & 0 \end{vmatrix}$$

Trigonal; 3

$$\begin{vmatrix} z_{11} & -z_{11} & 0 & z_{41} & z_{51} & -z_{22} \\ -z_{22} & z_{22} & 0 & z_{51} & -z_{41} & -z_{11} \\ z_{13} & z_{13} & z_{33} & 0 & 0 & 0 \end{vmatrix}$$

Trigonal; 32

$$\begin{vmatrix} z_{11} & -z_{11} & 0 & z_{41} & 0 & 0 \\ 0 & 0 & 0 & 0 & -z_{41} & -z_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Trigonal; 3m

$$\begin{vmatrix} 0 & 0 & 0 & 0 & z_{51} & -z_{22} \\ -z_{22} & z_{22} & 0 & z_{51} & 0 & 0 \\ z_{13} & z_{13} & z_{33} & 0 & 0 & 0 \end{vmatrix}$$

Hexagonal; $\bar{6}$

$$\begin{vmatrix} z_{11} & -z_{11} & 0 & 0 & 0 & -z_{22} \\ -z_{22} & z_{22} & 0 & 0 & 0 & -z_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Hexagonal; $\bar{6}$m2	z_{11}	$-z_{11}$	0	0	0	0
	0	0	0	0	0	$-z_{11}$
	0	0	0	0	0	0
Hexagonal; 6	0	0	0	z_{41}	z_{51}	0
	0	0	0	z_{51}	$-z_{41}$	0
	z_{13}	z_{13}	z_{33}	0	0	0
Hexagonal; 622	0	0	0	z_{41}	0	0
	0	0	0	0	$-z_{41}$	0
	0	0	0	0	0	0
Hexagonal; 6mm	0	0	0	0	z_{51}	0
	0	0	0	z_{51}	0	0
	z_{13}	z_{13}	z_{33}	0	0	0
Cubic; 23 and $\bar{4}$3m	0	0	0	z_{41}	0	0
	0	0	0	0	z_{41}	0
	0	0	0	0	0	z_{41}

The r tensor has similar terms.

The photoelastic constants are similar to the elastic constant tensors except that $m_{rs} \neq m_{sr}$ in general. However, for the tetragonal, trigonal, hexagonal and cubic systems, Pockels found that $m_{12} = m_{21}$. This follows from the transformation equations about the Z axis which is the n fold axes for these groups. For a rotation of an angle θ about Z, the direction cosines are

$$\begin{vmatrix} \ell_1 = \dfrac{\partial x_1'}{\partial x_1} = \cos\theta & m_1 = \dfrac{\partial x_1'}{\partial x_2} = \sin\theta & n_1 = \dfrac{\partial x_1'}{\partial x_3} = 0 \\ \ell_2 = \dfrac{\partial x_2'}{\partial x_1} = -\sin\theta & m_2 = \dfrac{\partial x_2'}{\partial x_2} = \cos\theta & n_2 = \dfrac{\partial x_2'}{\partial x_3} = 0 \\ \ell_3 = \dfrac{\partial x_3'}{\partial x_1} = 0 & m_3 = \dfrac{\partial x_3'}{\partial x_2} = 0 & n_3 = \dfrac{\partial x_3'}{\partial x_3} = 1 \end{vmatrix} \tag{92}$$

Transforming the two terms $m'_{1122} = m'_{12}$ and $m'_{2211} = m'_{21}$ by the tensor transformation equation

$$m_{ijk\ell} = \frac{\partial x'_i}{\partial x_m}\frac{\partial x'_j}{\partial x_n}\frac{\partial x'_k}{\partial x_o}\frac{\partial x'_\ell}{\partial x_p} m_{mnop} \tag{93}$$

we find, for these two coefficients,

$$\begin{aligned} m'_{12} &= (m_{11} + m_{22} - 4m_{66}) \sin^2\theta \cos^2\theta + 2(m_{62} - m_{16}) \\ &\quad \sin\theta \cos^3\theta + 2(m_{61} - m_{16}) \sin^3\theta \cos\theta + m_{12} \cos^4\theta + m_{21} \sin^4\theta \\ m'_{21} &= (m_{11} + m_{22} - 4m_{66}) \sin^2\theta \cos^2\theta + 2(m_{16} - m_{62}) \\ &\quad \sin^3\theta \cos\theta + 2(m_{26} - m_{61}) \sin\theta \cos^3\theta + m_{21} \cos^4\theta + m_{12} \sin^4\theta \end{aligned} \tag{94}$$

If $m'_{12} = m'_{21}$ for all angles of rotation we must have

$$m_{16} + m_{26} = m_{61} + m_{62}$$

For all the classes that $m_{12} = m_{21}$, either $m_{26} = -m_{16}$ and $m_{62} = -m_{61}$ or else $m_{16} = m_{26} = m_{61} = m_{62} = 0$.

Now, if Z is a four-fold axis, as it is in the tetragonal and cubic systems, then, for a 90° rotation, the value of m'_{12} or m'_{21} must repeat. From the first of (92) this means that

$$m_{12} = m_{21} \text{ and } m_{21} = m_{12}$$

For a trigonal or hexagonal system additional relations are obtained between m_{66} and m_{11}, m_{22} and m_{12} in the usual manner. Hence the photoelastic matrices become, for the various crystal classes,

(95)

Triclinic 36 Constant	m_{11}	m_{12}	m_{13}	m_{14}	m_{15}	m_{16}	The π tensor is entirely analogous
	m_{21}	m_{22}	m_{23}	m_{24}	m_{25}	m_{26}	
	m_{31}	m_{32}	m_{33}	m_{34}	m_{35}	m_{36}	
	m_{41}	m_{42}	m_{43}	m_{44}	m_{45}	m_{46}	
	m_{51}	m_{52}	m_{53}	m_{54}	m_{55}	m_{56}	
	m_{61}	m_{62}	m_{63}	m_{64}	m_{65}	m_{66}	
Monoclinic 20 Constants	m_{11}	m_{12}	m_{13}	0	m_{15}	0	The π tensor is entirely analogous
	m_{21}	m_{22}	m_{23}	0	m_{25}	0	
	m_{31}	m_{32}	m_{33}	0	m_{35}	0	
	0	0	0	m_{44}	0	m_{46}	
	m_{51}	m_{52}	m_{53}	0	m_{55}	0	
	0	0	0	m_{64}	0	m_{66}	

Orthorhombic 12 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & 0 & 0 & 0 \\ m_{21} & m_{22} & m_{23} & 0 & 0 & 0 \\ m_{31} & m_{32} & m_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & m_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & m_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & m_{66} \end{vmatrix}$$

The π tensor is entirely analogous

Tetragonal 4,4,4/*m* 9 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & 0 & 0 & m_{16} \\ m_{12} & m_{11} & m_{13} & 0 & 0 & -m_{16} \\ m_{31} & m_{31} & m_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & m_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & m_{44} & 0 \\ m_{61} & -m_{61} & 0 & 0 & 0 & m_{66} \end{vmatrix}$$

The π tensor is entirely analogous

Tetragonal 42*m*, 422 4*mm*, (4/*m*)*mm* 7 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & 0 & 0 & 0 \\ m_{12} & m_{11} & m_{13} & 0 & 0 & 0 \\ m_{31} & m_{31} & m_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & m_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & m_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & m_{66} \end{vmatrix}$$

The π tensor is entirely analogous

Trigonal 3,3 11 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & m_{14} & -m_{25} & 0 \\ m_{12} & m_{11} & m_{13} & -m_{14} & m_{25} & 0 \\ m_{31} & m_{31} & m_{33} & 0 & 0 & 0 \\ m_{41} & -m_{41} & 0 & m_{44} & m_{45} & m_{52} \\ -m_{52} & m_{52} & 0 & -m_{45} & m_{44} & m_{41} \\ 0 & 0 & 0 & m_{25} & m_{14} & \frac{m_{11}-m_{12}}{2} \end{vmatrix}$$

The π tensor is analogous except that $\pi_{46} = 2\pi_{52}$ $\pi_{56} = 2\pi_{41}$ $\pi_{66} = (\pi_{11} - \pi_{12})$

Trigonal 32,3*m* 3(2/*m*) 8 Constants

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & m_{14} & 0 & 0 \\ m_{12} & m_{11} & m_{13} & -m_{14} & 0 & 0 \\ m_{31} & m_{31} & m_{33} & 0 & 0 & 0 \\ m_{41} & -m_{41} & 0 & m_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & m_{44} & m_{41} \\ 0 & 0 & 0 & 0 & m_{14} & \frac{m_{11}-m_{12}}{2} \end{vmatrix}$$

The π tensor is analogous except that $\pi_{56} = 2\pi_{41}$ $\pi_{66} = \pi_{11} - \pi_{12}$

Hexagonal $6, 6m2, 6$ $622, 6/m;$ $6mm, \frac{6}{m} mm$ 6 Constants	m_{11}	m_{12}	m_{13}	0	0	0	The π tensor is analogous except that $\pi_{66} = \pi_{11} - \pi_{12}$
	m_{12}	m_{11}	m_{13}	0	0	0	
	m_{31}	m_{31}	m_{33}	0	0	0	
	0	0	0	m_{44}	0	0	
	0	0	0	0	m_{44}	0	
	0	0	0	0	0	$\frac{m_{11}-m_{12}}{2}$	
Cubic System $23, 432$ $\frac{2}{m}3, 43m, \frac{4}{m}3\frac{2}{m}$ 3 Constants	m_{11}	m_{12}	m_{12}	0	0	0	The π tensor is entirely analogous (95)
	m_{12}	m_{11}	m_{12}	0	0	0	
	m_{12}	m_{12}	m_{11}	0	0	0	
	0	0	0	m_{44}	0	0	
	0	0	0	0	m_{44}	0	
	0	0	0	0	0	m_{44}	
Isotropic Systems 2 Constants	m_{11}	m_{12}	m_{12}	0	0	0	The π tensor is analogous except that $\pi_{66} = \pi_{11} - \pi_{12}$
	m_{12}	m_{11}	m_{12}	0	0	0	
	m_{12}	m_{12}	m_{11}	0	0	0	
	0	0	0	$\frac{m_{11}-m_{12}}{2}$	0	0	
	0	0	0	0	$\frac{m_{11}-m_{12}}{2}$	0	
	0	0	0	0	0	$\frac{m_{11}-m_{12}}{2}$	

From measurement[7] on the photoelastic effects at high pressure for cubic crystals, it has become apparent that the second derivatives of equation (44) are not sufficient to represent the experimental results and derivatives up to the fourth power should be included. This extension, however, is not considered in the present paper.

VII. Photoelasticity in Isotropic Media

The photoelastic effect in isotropic solids has been used extensively in studying the stresses existing in machine parts and other pieces. For this purpose a plastic model cut in the shape of the original is used and is loaded in a similar manner to that of the machine part to be studied. Since stresses are applied, the π_i photoelastic constants are most useful. If we look along

[7] H. B. Maris, *Jour. Optical Society of Amer.*, Vol. 15, pp. 194–200, 1927.

the Z axis, the last of equations (79) shows that the birefringence is equal to

$$B_z = \frac{\mu^3}{2}\sqrt{(\beta_1 + \Delta_1 - \beta_2 - \Delta_2)^2 + 4(\Delta_6)^2} \tag{96}$$

Since, for an isotropic substance $\beta_1 = \beta_2$, we have, after substituting the value of Δ_1 and Δ_2, with the appropriate photoelastic constants from equation (95), (last tensor):

$$B_z = \frac{\mu^3}{2}(\pi_{11} - \pi_{12})\sqrt{(T_1 - T_2)^2 + 4T_6^2} \tag{97}$$

If we transform to axes rotated by an angle θ about Z, the values of T'_{11} and T'_{22} are given by

$$\begin{aligned} T'_{11} &= \cos^2\theta T_1 + 2\sin\theta\cos\theta T_6 + \sin^2\theta T_2 \\ T'_{22} &= \sin^2\theta T_1 - 2\sin\theta\cos\theta T_6 + \cos^2\theta T_2 \end{aligned} \tag{98}$$

If, now, we choose the angle θ so that T'_{11} is a maximum, we find

$$\tan 2\theta = \frac{+2T_6}{T_1 - T_2} \tag{99}$$

Inserting this value of $\tan 2\theta$ in (98) we find

$$\begin{aligned} T'_1 &= \frac{T_1 + T_2}{2} + \tfrac{1}{2}\sqrt{(T_1 - T_2)^2 + 4T_6^2} \\ T'_2 &= \frac{T_1 + T_2}{2} - \tfrac{1}{2}\sqrt{(T_1 - T_2)^2 + 4T_6^2} \end{aligned} \tag{100}$$

and, hence,

$$T'_1 - T'_2 = \sqrt{(T_1 - T_2)^2 + 4T_6^2} \tag{101}$$

Hence the birefringence obtained in stressing a material is proportional to the difference in the principal stresses. By observing the isoclinic lines of a photoelastic picture, methods[8] are available for determining the stresses in a model. A photograph[9] of a stressed disk is shown by Fig. 3. The high concentration of lines near the surface shows that the shearing stress is very high at these points. By counting the number of lines from the edge and knowing the stress optical constant, the stress can be calculated at any point.

If we apply a single stress T_1, the birefringence is given by the equation

$$B_z = \frac{\mu^3}{2}(\pi_{11} - \pi_{12})T_1 \tag{102}$$

[8] See Photoelasticity, Coker and Filon, Cambridge University Press, 1931.
[9] This photograph was taken by T. F. Osmer.

Instead of using the constants π_{11} and π_{12} it is customary to use a single constant C given by

$$B = \mu_e - \mu_o = r = CT \tag{103}$$

where the constant C is called the relative stress optical constant and r the retardation. The dimensions of C are the reciprocal of a stress and are

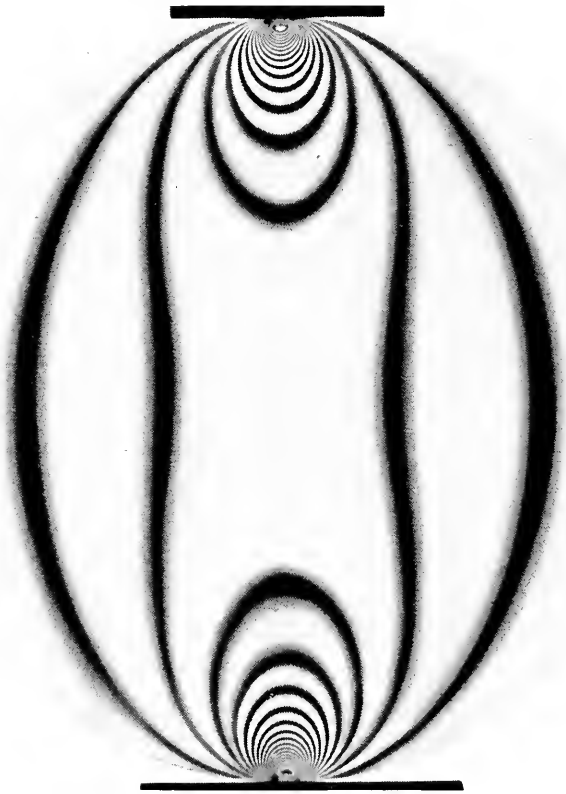

Fig. 3—Photoelastic picture of a disk in compression.

measured in cm^2 per dyne. A convenient unit for most purposes is one of 10^{-13} cm^2/dyne; if this is used, the stress optical coefficients of most glasses are from 1 to 10 and most plastics are from 10 to 100. This unit so defined has been called the "Brewster". In terms of the Brewster, the retardation is

$$r = CTd \tag{104}$$

If C is measured in Brewsters, d in millimeters and T in bars (10^6 dynes/cm^2) then r, as given by the formula, is expressed in angstrom units.

Traveling-Wave Tubes

By J. R. PIERCE

[SECOND INSTALLMENT]

CHAPTER IV

FILTER-TYPE CIRCUITS

Synopsis of Chapter

ASIDE FROM HELICES, the circuits most commonly used in traveling-wave tubes are iterated or filter-type circuits, composed of linear arrays of coupled resonant slots or cavities.

Sometimes the geometry of such structures is simple enough so that an approximate field solution can be obtained. In other cases, the behavior of the circuits can be inferred by considering the behavior of lumped-circuit analogues, and the behavior of the circuits with frequency can be expressed with varying degrees of approximation in terms of parameters which can be computed or experimentally evaluated.

In this chapter the field approach will be illustrated for some very simple circuits, and examples of lumped-circuit analogues of other circuits will be given. The intent is to present methods of analyzing circuits rather than particular numerical results, for there are so many possible configurations that a comprehensive treatment would constitute a book in itself.

Readers interested in a wider and more exact treatment of field solutions are referred to the literature.[1,2]

The circuit of Fig. 4.1 is one which can be treated by field methods. This "corrugated waveguide" type of circuit was first brought to the writer's attention by C. C. Cutler. It is composed of a series of parallel equally spaced thin fins of height h projecting normal to a conducting plane. The case treated is that of propagation of a transverse magnetic wave, the magnetic field being parallel to the length of the fins. It is assumed that the spacing ℓ is small compared with a wavelength. In Fig. 4.2, βh is plotted vs. $\beta_0 h$. Here β is the phase constant and $\beta_0 = \omega/c$ is a phase constant corresponding to the velocity of light.

[1] E. L. Chu and W. W. Hansen, "The Theory of Disk-Loaded Wave Guides," *Journal of Applied Physics*, Vol. 18, pp. 999–1008, Nov. 1947.

[2] L. Brillouin, "Wave Guides for Slow Waves," *Journal of Applied Physics*, Vol. 19, pp. 1023–1041, Nov. 1948.

For small values of $\beta_0 h$, that is, at low frequencies, very nearly $\beta = \beta_0$; that is, the phase velocity is very near to the velocity of light. The field decays slowly away from the circuit. The longitudinal electric field is small compared with the transverse electric field. In fact, as the frequency approaches zero, the wave approaches a transverse electromagnetic wave traveling with the speed of light.

At high frequencies the wave falls off rapidly away from the circuit, and the transverse and longitudinal components of electric field are almost equal. The wave travels very slowly. As the wavelength gets so short that the spacing ℓ approaches a half wavelength ($\beta\ell = \pi$) the simple analysis given is no longer valid. Actually, $\beta\ell = \pi$ specifies a cutoff frequency; the circuit behaves as a lowpass filter.

Figure 4.3 shows two opposed sets of fins such as those of Fig. 4.1. Such a circuit propagates two modes, a transverse mode for which the longitudinal electric field is zero at the plane of symmetry and a longitudinal mode for which the transverse electric field is zero at the plane of symmetry.

At low frequencies, the longitudinal mode corresponds to the wave on a loaded transmission line. The fins increase the capacitance between the conducting planes to which they are attached but they do not decrease the inductance. Figure 4.6 shows βh vs. $\beta_0 h$ for several ratios of fin height, h, to half-separation, d. The greater is h/d, the slower is the wave (the larger is β/β_0).

The longitudinal mode is like a transverse magnetic waveguide mode; it propagates only at frequencies above a cutoff frequency, which increases as h/d is increased. Figure 4.7 shows βh vs. $\beta_0 h = (\omega/c)h$ for several values of h/d. The cutoff, for which $\beta\ell = \pi$, occurs for a value of $\beta_0 h$ less than $\pi/2$. Thus, we see that the longitudinal mode has a band pass characteristic. The behavior of the longitudinal mode is similar to that of a longitudinal mode of the washer-loaded waveguide shown in Fig. 4.8. The circuit of Fig. 4.8 has been proposed for use in traveling-wave tubes.

The transverse mode of the circuit of Fig. 4.3 can also exist in a circuit consisting of strips such as those of Fig. 4.1 and an opposed conducting plane, as shown in Fig. 4.5. This circuit is analogous in behavior to the disk-on-rod circuit of Fig. 4.9. The circuit of Fig. 4.5 may be thought of as a loaded parallel strip line. That of Fig. 4.9 may be thought of as a loaded coaxial line.

Wave-analysis makes it possible to evaluate fairly accurately the transmission properties of a few simple structures. However, iterated or repeating structures have certain properties in common: the properties of filter networks.

For instance, a mode of propagation of the loaded waveguide of Fig. 4.10 or of the series of coupled resonators of Fig. 4.11 can be represented accurately at a single frequency by the ladder networks of Fig. 4.12. Further,

if suitable lumped-admittance networks are used to represent the admittances B_1 and B_2, the frequency-dependent behavior of the structures of Figs. 4.10 and 4.11 can be approximated.

It is, for instance, convenient to represent the shunt admittances B_2 and the series admittances B_1 in terms of a "longitudinal" admittance B_L and a "transverse" admittance B_T. B_L and B_T are admittances of shunt resonant circuits, as shown in Fig. 4.15, where their relation to B_1 and B_2 and approximate expressions for their frequency dependence are given. The resonant frequencies of B_L and B_T, that is, ω_L and ω_T, have simple physical meanings. Thus, in Fig. 4.10, ω_L is the frequency corresponding to equal and opposite voltages across successive slots, that is, the π mode frequency. ω_T is the frequency corresponding to zero slot voltage and no phase change along the filter, that is, the zero mode frequency.

If ω_L is greater than ω_T, the phase characteristic of this lumped-circuit analogue is as shown in Fig. 4.17. The phase shift is zero at the lower cutoff frequency ω_T and rises to π at the upper cutoff frequency ω_L. If ω_T is greater than ω_L, the phase shift starts at $-\pi$ at the lower cutoff frequency ω_L and rises to zero at the upper cutoff frequency ω_T, as shown in Fig. 4.19. In this case the phase velocity is negative. Figure 4.20 shows a measure of $(E^2/\beta^2 P)$ plotted vs. ω for $\omega_L > \omega_T$. This impedance parameter is zero at ω_T and rises to infinity at ω_L.

The structure of Fig. 4.11 can be given a lumped-circuit equivalent in a similar manner. In this case the representation should be quite accurate. We find that ω_L is always greater than ω_T and that one universal phase curve, shown in Fig. 4.27, applies. A curve giving a measure of $(E^2/\beta^2 P)$ vs. frequency is shown in Fig. 4.28. In this case the impedance parameter goes to infinity at both cutoff frequencies.

The electric field associated with iterated structures does not vary sinusoidally with distance but it can be analyzed into sinusoidal components. The electron stream will interact strongly with the circuit only if the electron velocity is nearly equal to the phase velocity of one of these field components. If θ is the phase shift per section and L is the section length, the phase constant β_m of a typical component is

$$\beta_m = (\theta + 2m\pi)/L$$

where m is a positive or negative integer. The field component for which $m = 0$ is called the fundamental; for other values of m the components are called *spatial harmonics*. Some of these components have negative phase velocities and some have positive phase velocities.

The peak field strength of any field component may be expressed

$$E = -M(V/L)$$

Here V is the peak gap voltage, L is the section spacing and M is a function of β (or β_m) and of various dimensions. For the electrode systems of Figs.

4.29, 4.30, 4.31 and 4.32 M is given by (4.69), (4.71), (4.72) and (4.73), respectively.

The factor M may be indifferently regarded as a factor by which we multiply the a-c beam current to give the induced current at the gap, or, as a factor by which we multiply the gap voltage in obtaining the field. We can go further, evaluate $E^2/\beta^2 P$ in terms of gap voltage, and use $M^2 I_0$ as the effective current, or we can use the current I_0 and take the effective field in the impedance parameter as

$$E^2 = M^2(V/\ell)^2$$

It is sometimes desirable to make use of a spatial harmonic ($m \neq 0$) instead of a fundamental, usually to (1) allow a greater resonator spacing (2) to obtain a positive phase velocity when the fundamental has a negative phase velocity (3) to obtain a phase curve for which the phase angle is nearly a constant times frequency; that is, a phase curve for which the group velocity does not change much with frequency and hence can be matched by the electron velocity over a considerable frequency range. Figure 4.33 shows how $\theta + 2\pi$ (the phase shift per section for $m = 1$) can be nearly a constant times ω even when θ is not.

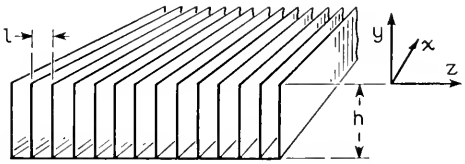

Fig. 4.1—A corrugated or finned circuit with filter-like properties.

4.1 Field Solutions

An approximate field analysis will be made for two very simple two-dimensional structures. The first of these, which is shown in Fig. 4.1, is empty space for $y > 1$ and consists of very thin conducting partitions in the y direction from $y = 0$ to $y = -h$; the partitions are connected together by a conductor in the z direction at $y = -h$. These conducting partitions are spaced a distance ℓ apart in the z direction. The structure is assumed to extend infinitely in the $+x$ and $-x$ directions.

In our analysis we will initially assume that the wavelength of the propagated wave is long compared with ℓ. In this case, the effect of the partitions is to prevent the existence of any y component of electric field below the z axis, and the conductor at $y = -h$ makes the z component of electric field zero at $y = -z$.

In some perfectly conducting structures the waves propagated are either transverse electric (no electric field component in the direction of propagation, that is, z direction) or transverse magnetic (no magnetic field com-

ponent in the z direction). We find that for the structure under consideration there is a transverse magnetic solution. We can take it either on the basis of other experience or as a result of having solved the problem that the correct form for the x component of magnetic field for $y > 0$ is

$$H_x = H_0 e^{(-\gamma y - j\beta z)} \tag{4.1}$$

Expressing the electric field in terms of the curl of the magnetic field, we have

$$\begin{aligned} j\omega\epsilon E_x &= \frac{\partial Hz}{\partial y} - \frac{\partial Hy}{\partial z} = 0 \\ j\omega\epsilon E_y &= \frac{\partial Hx}{dz} - \frac{\partial Hz}{\partial x} \end{aligned} \tag{4.2}$$

$$E_y = -\frac{\beta}{\omega\epsilon} H_0 e^{(-\gamma y - j\beta z)} \tag{4.3}$$

$$j\omega\epsilon E_z = \frac{\partial Hy}{\partial x} - \frac{\partial Hx}{\partial y} \tag{4.4}$$

$$E_z = -j\frac{\gamma}{\omega\epsilon} H_0 e^{(-\gamma y - j\beta z)} \tag{4.5}$$

We can in turn express H_x in terms of E_y and E_z

$$-j\omega\mu H_x = \frac{\partial Ez}{\partial y} - \frac{\partial Ey}{\partial z} \tag{4.6}$$

This leads to the relation

$$\beta^2 - \gamma^2 = \omega^2\mu\epsilon \tag{4.7}$$

Now, $1/\sqrt{\mu\epsilon}$ is the velocity of light, and ω divided by the velocity of light has been called β_0, so that

$$\beta^2 - \gamma^2 = \beta_0^2 \tag{4.8}$$

Between the partitions, the field does not vary in the z direction. In any space between from $y = 0$ to $y = -h$, the appropriate form for the magnetic field is

$$H_x = H_0 \frac{\cos \beta_0(y + h)}{\cos \beta_0 h} \tag{4.9}$$

From this we obtain by means of (4.4)

$$E_z = -\frac{j\beta_0}{\omega\epsilon} H_0 \frac{\sin \beta_0(y + h)}{\cos \beta_0 h} \tag{4.10}$$

Application of (4.6) shows that this is correct.

Now, at $y = 0$ we have just above the boundary

$$E_z = -j \frac{\gamma}{\omega\epsilon} H_0 e^{-j\beta z} \tag{4.11}$$

The fields in the particular slot just below the boundary will be in phase with these (we specify this by adding a factor exp $-j\beta z$ to 4.10) and hence will be

$$E_z = -\frac{j\beta_0}{\omega\epsilon} H_0 e^{-j\beta z} \tan \beta_0 h \tag{4.12}$$

From (4.11) and (4.12) we see that we must have

$$\beta_0 h \tan \beta_0 h = \gamma h \tag{4.13}$$

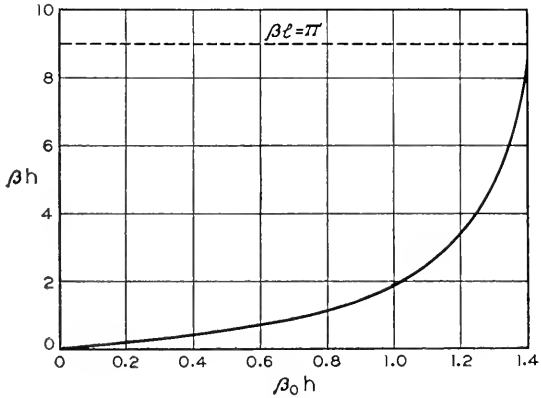

Fig. 4.2—The approximate variation of the phase constant β with frequency (proportional to $\beta_0 h$) for the circuit of Fig. 4.1. The curve is in error as $\beta\ell$ approaches π, and there is a cutoff at $\beta\ell = \pi$.

Using (4.8), we obtain

$$\beta h = \frac{\pm\beta_0 h}{\cos \beta_0 h} \tag{4.14}$$

In Fig. 4.2, βh has been plotted vs $\beta_0 h$, which is, of course, proportional to frequency. This curve starts out as a straight line, $\beta = \beta_0$; that is, for low frequencies the speed is the speed of light. At low frequencies the field falls off slowly in the y direction, and as the frequency approaches zero we have essentially a plane electromagnetic wave. At higher frequencies, $\beta > \beta_0$, that is, the wave travels with less than the speed of light, and the field falls off rapidly in the y direction. According to (4.14), β goes to infinity at $\beta_0 h = \pi/2$.

As a matter of fact, the match between the fields assumed above and below the boundary becomes increasingly bad as $\beta\ell$ becomes larger. The most rapid

alteration we can have below the boundary is one in which fields in alternate spaces follow a $+, -, +, -$ pattern. Thus, the rapid variations of field above the boundary predicted by (4.14) for values of $\beta_0 h$ which make $\beta \ell$ greater than π cannot be matched below the boundary. The frequency at which $\beta \ell = \pi$ constitutes the cutoff frequency of the structure regarded as a filter. There is another pass band in the region $\pi < \beta_0 h < 3\pi/2$, in which the ratio of E to H below the boundary has the same sign as the ratio of E to H above the boundary.

A more elaborate matching of fields would show that our expression is considerably in error near cutoff. This matter will not be pursued here; the behavior of filters near cutoff will be considered in connection with lumped circuit representations.

We can obtain the complex power flow P by integrating the Poynting vector over a plane normal to the z direction in the region $y > 0$. Let us consider the power flow over a depth W normal to the plane of the paper. Then

$$P = \frac{1}{2} \int_0^\infty \int_0^W (E_x H_y^* - E_y H_x^*) \, dx \, dy \tag{4.15}$$

Using (4.1) and (4.3), we obtain

$$\begin{aligned} P &= \frac{W}{2} \int_0^\infty \frac{\beta H_0^2}{\omega \epsilon} e^{-2\gamma y} \, dy \\ P &= \frac{1}{4} \frac{H_0^2 \beta W}{\omega \epsilon \gamma} \end{aligned} \tag{4.16}$$

We will express this in terms of E the magnitude of the z component of the field at $y = 0$, which, according to (4.5), is

$$E = \frac{\gamma}{\omega \epsilon} H_0 \tag{4.17}$$

We will also note that

$$\begin{aligned} \omega \epsilon &= \omega \sqrt{\mu \epsilon} / \sqrt{\mu / \epsilon} \\ &= (\omega / c) / \sqrt{\mu / \epsilon} = \beta_0 / \sqrt{\mu / \epsilon} \end{aligned} \tag{4.18}$$

and that

$$\sqrt{\mu / \epsilon} = 377 \text{ ohms} \tag{4.19}$$

By using (4.17)–(4.18) in connection with (4.16), we obtain

$$E^2 / \beta^2 P = (4 / \beta_0 W)(\gamma / \beta)^3 \sqrt{\mu / \epsilon} \tag{4.20}$$

We notice that this impedance is very small for low frequencies, at which

the velocity of the wave is high, and the field extends far in the y direction and becomes higher at high frequencies, where the velocity is low and the field falls off rapidly.

We will next consider a symmetrical array of two opposed sets of slots (Fig. 4.3) similar to that shown in Fig. 4.1. Two modes of propagation will be of interest. In one the field is symmetrical about the axis of physical symmetry, and in the other the fields at positions of physical symmetry are equal and opposite.

In writing the equations, we need consider only half of the circuit. It is convenient to take the z axis along the boundary, as shown in Fig. 4.4.

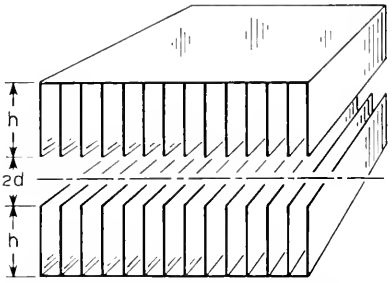

Fig. 4.3—A double finned structure which will support a transverse mode (no longitudinal electric field on axis) and a longitudinal mode (no transverse electric field on axis).

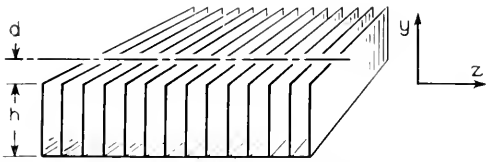

Fig. 4.4—The coordinates used in connection with the circuit of Fig. 4.3.

This puts the axis of symmetry at $y = +d$, and the slots extend from $y = 0$ to $y = -h$.

For negative values of y, (4.9), (4.10), (4.12) hold.

Let us first consider the case in which the fields above are opposite to the fields below. This also corresponds to waves in a series of slots opposite a conducting plane, as shown in Fig. 4.5. In this case the appropriate form of the magnetic field above the boundary is

$$H_x = H_0 \frac{\cosh \gamma(d - y)}{\cosh \gamma d} e^{-j\beta z} \tag{4.21}$$

From Maxwell's equations we then find

$$E_y = -\frac{\beta}{\omega\epsilon} H_0 \frac{\cosh \gamma(d - y)}{\cosh \gamma d} e^{-j\beta z} \tag{4.22}$$

$$E_z = -j \frac{\gamma}{\omega\epsilon} H_0 \frac{\sinh \gamma(d - y)}{\cosh \gamma d} e^{-j\beta z} \tag{4.23}$$

$$\beta_0^2 = \beta^2 - \gamma^2 \tag{4.24}$$

At $y = 0$ we have from (4.23) and (4.12)

$$E_z = -j \frac{\gamma}{\omega\epsilon} H_0 e^{-j\beta z} \tanh \gamma d \tag{4.25}$$

$$E_z = -j \frac{\beta_0}{\omega\epsilon} H_0 e^{-j\beta z} \tan \beta_0 h \tag{4.12}$$

Hence, we must have

$$\gamma h \tanh ((d/h)\gamma h) = \beta_0 h \tan \beta_0 h \tag{4.26}$$

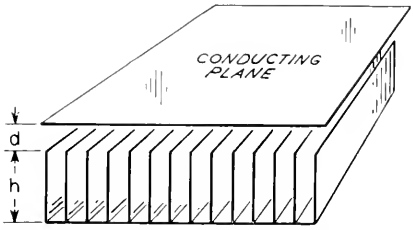

Fig. 4.5—The transverse mode of the circuit of Fig. 4.3 exists in this circuit also.

Here we have added parameter, (d/h). For any value of d/h, we can obtain γh vs $\beta_0 h$; and we can obtain βh in terms of γh by means of 4.24

$$\beta h = ((\gamma h)^2 + (\beta_0 h)^2)^{1/2} \tag{4.27}$$

We see that for small values of $\beta_0 h$ (low frequencies)

$$\gamma^2 = (h/d)\, \beta_0^2 \tag{4.28}$$

$$\beta = \beta_0 \left(\frac{h + d}{d}\right)^{1/2} \tag{4.29}$$

If we examine Fig. 4.5, to which this applies, we find (4.28) easy to explain. At low frequencies, the magnetic field is essentially constant from $y = d$ to $y = -h$, and hence the inductance is proportional to the height $h + d$. The electric field will, however, extend only from $y = 0$ to $y = d$; hence the capacitance is proportional to $1/d$. The phase constant is proportional to $\sqrt{LC}$, and hence (4.29). At higher frequencies the electric and magnetic fields vary with y and (4.29) does not hold.

We see that (4.26) predicts infinite values of γ for $\beta h = \pi/2$. As in the previous cases, cutoff occurs at $\beta\ell = \pi$.

As an example of the phase characteristic of the circuit, βh from (4.26) and (4.27) is plotted vs $\beta_0 h$ for $h/d = 0$, 10, 100 in Fig. 4.6. The curve for $h/d = 0$ is of course the same as Fig. 4.2.

If we integrate Poynting's vector from $y = 0$ to $y = d$ and for a distance W in the x direction, and multiply by 2 to take the power flow in the other half of the circuit into account, we obtain

$$E^2/\beta^2 P = (2/\beta_0 W)(\gamma/\beta)^3 \left(\frac{\sinh^2 \gamma d}{\sinh \gamma d \cosh \gamma d + \gamma d} \right) \sqrt{\mu/\epsilon} \quad (4.30)$$

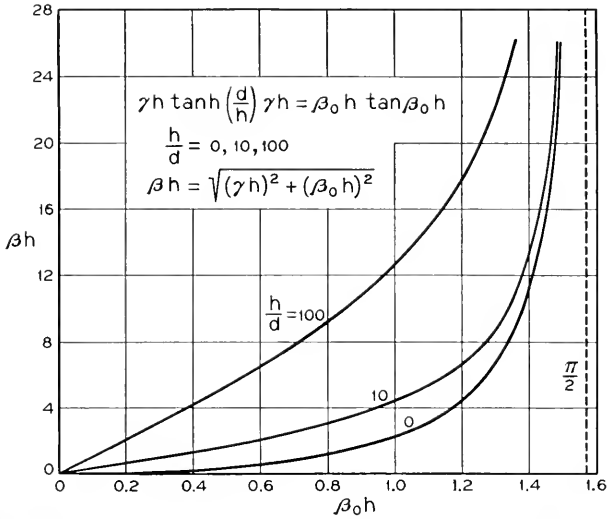

Fig. 4.6—The variation of β with frequency (proportional to $\beta_0 h$) for the transverse mode of the circuit of Fig. 4.3. Again, the curves are in error near the cutoff at $\beta \ell = \pi$.

At very low frequencies, at which (4.28) and (4.29) hold, we have

$$\begin{aligned} E^2/\beta^2 P &= (\gamma^4/\beta_0 \beta^3)(d/W) \sqrt{\mu/\epsilon} \\ E^2/\beta^2 P &= (h/d)^{1/2} (1 + d/h)^{3/2} (d/W) \sqrt{\mu/\epsilon} \end{aligned} \quad (4.31)$$

At high frequencies, for which γd is large, (4.30) approaches $\frac{1}{2}$ of the value given by (4.20). There is twice as much power because there are two halves to the circuit.

Let us now consider the case in which the field is symmetrical and E_z does not go to zero on the axis. In this case the appropriate field for $y > 0$ is

$$H_x = H_0 \frac{\sinh \gamma(d - y)}{\sinh \gamma d} e^{-j\beta z} \quad (4.32)$$

Proceeding as before, we find

$$\frac{\gamma h}{\tanh\left((d/h)\,\gamma h\right)} = \beta_0 h \tan \beta_0 h \tag{4.33}$$

We see that, in this case, for small values of γh we have

$$\beta_0 h \tanh \beta_0 h = h/d \tag{4.33a}$$

There is no transmission at all for frequencies below that specified by (4.33). As the frequency is increased above this lower cutoff frequency, γh and hence βh increase, and approach infinity at $\beta_0 h = \pi/2$. Actually, of course, the upper cutoff occurs at $\beta \ell = \pi$. In Fig. 4.7 βh is plotted vs $\beta_0 h$ for $h/d = 0$,

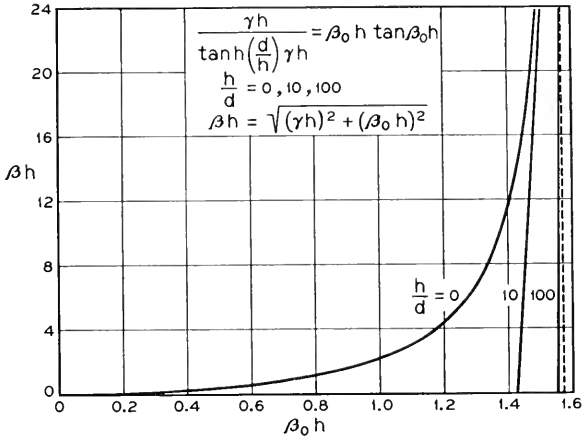

Fig. 4.7—The variation of β with frequency (proportional to $\beta_0 h$) for the longitudinal mode of the circuit of Fig. 4.3. This mode has a band pass characteristic; the band narrows as the opening of width 2d is made small compared with the fin height. Again, the curves are in error near the upper cutoff at $\beta \ell = \pi$.

10, 100. This illustrates how the band is narrowed as the opening between the slots is decreased.

By the means used before we obtain

$$E^2/\beta^2 P = (2/\beta_0 W)(\gamma/\beta)^3 \left(\frac{\cosh^2 \gamma d}{\sinh \gamma d \cosh \gamma d - \gamma d}\right) \sqrt{\mu/\epsilon} \tag{4.34}$$

We see that this goes to infinity at $\gamma d = 0$. For large values of γd it becomes the same as (4.30).

4.2 Practical Circuits

Circuits have been proposed or used in traveling-wave tubes which bear a close resemblance to those of Figs. 4.1, 4.3, 4.5 and which have very similar

properties[3]. Thus Field[4] describes an apertured disk structure (Fig. 4.8) which has band-pass properties very similar to the symmetrical mode of the circuit of Fig. 4.3. In this case there is no mode similar to the other mode, with equal and opposite fields in the two halves. Field also shows a disk-on-rod structure (Fig. 4.9) and describes a tube using it. This structure has low-

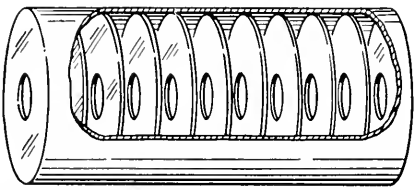

Fig. 4.8—This loaded waveguide circuit has band-pass properties similar to those of Fig. 4.7.

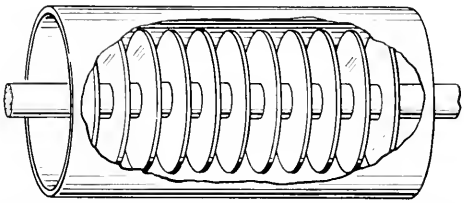

Fig. 4.9—This disk-on-rod circuit has properties similar to those of Fig. 4.6.

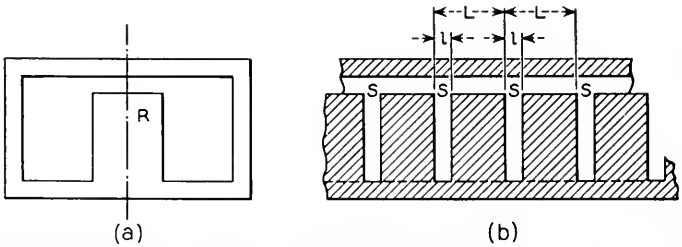

Fig. 4.10—A circuit consisting of a ridged waveguide with transverse slots or resonators in the ridge.

pass properties very similar to those of the circuit of Fig. 4.5, which are illustrated in Fig. 4.6.

Figure 4.10 shows a somewhat more complicated circuit. Here we have a rectangular waveguide, shown end on in *a* of Fig. 4.10, loaded by a longitudinal ridged portion *R*. In *b* of Fig. 4.10 we have a longitudinal cross sec-

[3] F. B. Llewellyn, *U. S. Patents* 2,367,295 and 2,395,560.

[4] Lester M. Field, "Some Slow-Wave Structures for Traveling-Wave Tubes," *Proc. I.R.E.*, Vol. 37, pp. 34–40, Jan. 1949.

tion, showing regularly spaced slots S cut in the ridge R. The slots S may be thought of as resonators.

Figure 4.11 shows in cross section a circuit made of a number of axially symmetrical reentrant resonators R, coupled by small holes H which act as inductive irises.

It would be very difficult to apply Maxwell's equations directly in deducing the performance of the structures shown in Figs. 4.10 and 4.11. Moreover, it is apparent that we can radically change the performance of

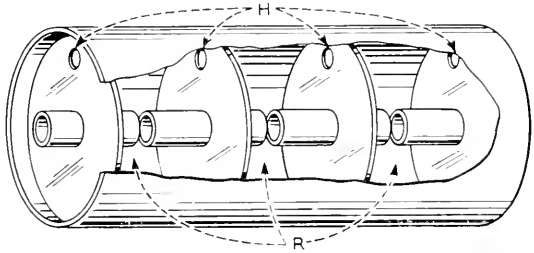

Fig. 4.11—A circuit consisting of a number of resonators inductively coupled by means of holes.

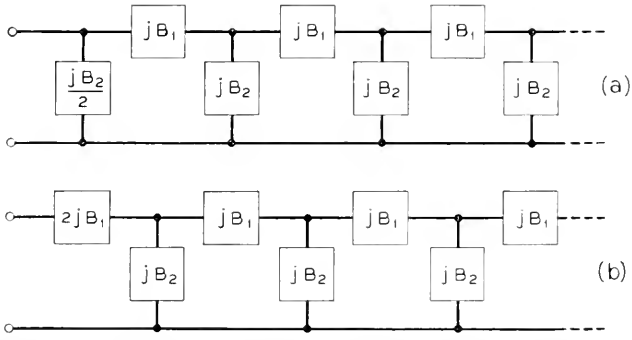

Fig. 4.12—Ladder networks terminated in π (above) and T (below) half sections. Such networks can be used in analyzing the behavior of circuits such as those of Figs. 4.10 and 4.11.

such structures by minor physical alterations as, by changing the iris size, or by using resonant irises in the circuit of Fig. 4.11, for instance.

As a matter of fact, it is not necessary to solve Maxwell's equations afresh each time in order to understand the general properties of these and other circuits.

4.3 Lumped Iterated Analogues

Consider the ladders of lossless admittances or susceptances shown in Fig. 4.12. Susceptances rather than reactances have been chosen because the

elements we shall most often encounter are shunt resonant near the frequencies considered; their susceptance is near zero and changing slowly but their reactance is near infinity.

If these ladders are continued endlessly to the right (or terminated in a reflectionless manner) and if a signal is impressed on the left-hand end, the voltages, currents and fields at corresponding points in successive sections will be in the ratio exp(-Γ) so that we can write the voltages,

$$V_n = V_0 e^{-n\Gamma} \tag{4.35}$$

If the admittances Y_1 and Y_2 are pure susceptances (lossless reactors), Γ is either purely real (an exponential decay with distance) or purely imaginary (a pass band). In this case Γ is usually replaced by $j\beta$. In order to avoid confusion of notation, we will use $j\theta$ instead, and write for the lossless case in the pass band

$$V_n = V_0 e^{-jn\theta} \tag{4.35a}$$

Thus, θ is the phase lag in radians in going from one section to the next. In terms of the susceptances,*

$$\cos\theta = 1 + B_2/2B_1 \tag{4.36}$$

We will henceforward assume that all elements are lossless.

Two characteristic impedances are associated with such iterated networks. If the network starts with a shunt susceptance $B_1/2$, as in *a* of Fig. 4.12, then we see the mid-shunt characteristic impedance K_π

$$K_\pi = 2(-B_2(B_2 + 4B_1))^{-1/2} \tag{4.37}$$

If the network starts with a series susceptance $2B_1$ we see the mid-series characteristic impedance K_T

$$K_T = \pm(1/2B_1)(-B_2 + 4B_1)/B_2)^{1/2} \tag{4.38}$$

Here the sign is chosen to make the impedance positive in the pass band.

When such networks are used as circuits for a traveling-wave tube, the voltage acting on the electron stream may be the voltage across B_2 or the voltage across B_1 or the voltage across some capacitive element of B_2 or B_1. We will wish to relate this peak voltage V to the power flow P. If the voltage across B_2 acts on the electron stream

$$V^2/P = 2K_\pi \tag{4.39}$$

If the voltage across Y_1 acts on the electron stream

$$V = I/jB_1$$

* The reader can work such relations out or look them up in a variety of books or handbooks. They are in Schelkunoff's Electromagnetic Waves.

where I is the current in B_1

$$P = |I^2| K_T/2$$

and hence

$$V^2/P = 2/B_1^2 K_T \tag{4.40}$$

$$V^2/P = -4(B_2/B_1)(-B_2(B_2 + 4B_1))^{-1/2} \tag{4.41}$$

$$V^2/P = -2(B_2/B_1)K_\pi \tag{4.42}$$

Here the sign has been chosen so as to make V^2/P positive in the pass band.

Let us now consider as an example the structure of Fig. 4.10. We see that two sorts of resonance are possible. First, if all the slots are shorted, or if no voltage appears between them, we can have a resonance in which the field between the top of the ridge R and the top of the waveguide is constant

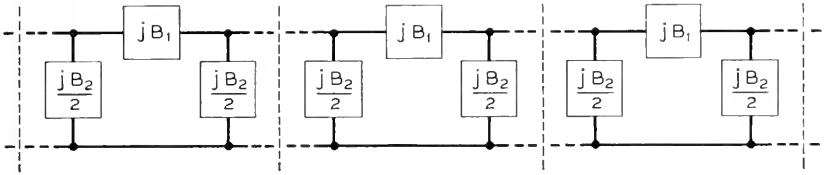

Fig. 4.13—A ladder network broken up into π sections.

all along the length, and corresponds to the cutoff frequency of the ridged waveguide. There are no longitudinal currents (or only small ones near the slots S) and hence there is no voltage across the slots and their admittance (the slot depth, for instance) does not affect the frequency of this resonance. Looking at Fig. 4.12, we see that this corresponds to a condition in which all shunt elements are open, or $B_2 = 0$. We will call the frequency of this resonance ω_T, the T standing for transverse.

There is another simple resonance possible; that in which the fields across successive slots are equal and opposite. Looking at Fig. 4.12, we see that this means that equal currents flow into each shunt element from the two series elements which are connected to it. We could, in fact, divide the network up into unconnected π sections, associating with each series element of susceptance B_1 half of the susceptance of a shunt element, that is, $B_2/2$, at each end, as shown in Fig. 4.13, without affecting the frequency of this resonance. This resonance, then, occurs at the frequency ω_L (L for longitudinal) at which

$$B_1 + B_2/4 = 0. \tag{4.43}$$

We have seen that the transverse resonant frequency, ω_T, has a clear meaning in connection with the structure of Fig. 4.10; it is (except for small

errors due to stray fields near the slots) the cutoff frequency of the waveguide without slots. Does the longitudinal frequency ω_L have a simple meaning?

Suppose we make a model of one section of the structure, as shown in Fig. 4.14. Comparing this with *b* of Fig. 4.10, we see that we have included the section of the ridged portion between two slots, and one half of a slot at each end, and closed the ends off with conducting plates *C*. The resonant frequency of this model is ω_L, the longitudinal resonant frequency defined above.

We will thus liken the structure of Fig. 4.10 to the filter network of Fig.

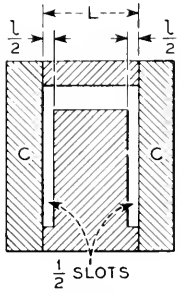

Fig. 4.14—A section which will have a resonant frequency corresponding to that for π radians phase shift per section in the circuit of Fig. 4.10.

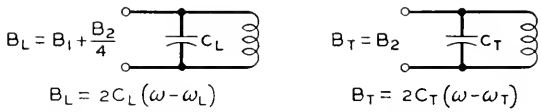

Fig. 4.15—The approximate variation with frequency (over a narrow band) of the longitudinal (B_L) transverse (B_T) susceptances of a filter network.

4.12, and express the susceptances B_1 and B_2 in terms of two susceptances B_T and B_L associated with the transverse and longitudinal resonances and defined below

$$B_T = B_2 \tag{4.44}$$

$$B_L = B_1 + B_2/4 \tag{4.45}$$

At the transverse resonant frequency ω_T, $B_T = 0$, and at the longitudinal resonant frequency ω_L, $B_L = 0$. So far, the lumped-circuit representation of the structure of Fig. 4.14 can be considered exact in the sense that at any frequency we can assign values to B_T and B_L which will give the correct values for θ and for V^2/P for the voltage across either the shunt or the series elements (whichever we are interested in).

We will go further and assume that near resonances these values of B_T and B_L behave like the admittances of shunt resonant circuits, as indicated in Fig. 4.15. Certainly we are right by our definition in saying that $B_T = 0$ at ω_T, and $B_L = 0$ at ω_L. We will assume near these frequencies a linear variation of B_T and B_L with frequency, which is very nearly true for shunt resonant circuits near resonance*

$$B_T = 2C_T(\omega - \omega_T) \tag{4.46}$$

$$B_L = 2C_L(\omega - \omega_L) \tag{4.47}$$

Here C_T can mean twice the peak stored electric energy per section length for unit peak voltage between the top of the guide and the top of the ridge R when the structure resonates in the transverse mode, and C_L can mean twice the stored energy per section length L for unit peak voltage across the top

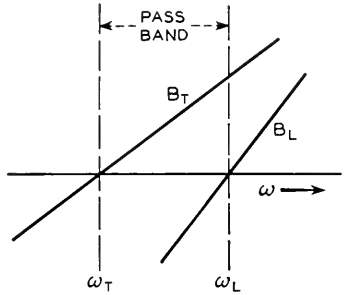

Fig. 4.16—Longitudinal and transverse susceptances which give zero radians phase shift at the lower cutoff ($\omega = \omega_T$) and π radians phase shift at the upper cutoff ($\omega = \omega_L$).

of the slot when the structure resonates in the longitudinal mode.

In terms of B_T and B_L, expression (4.36) for the phase angle θ becomes

$$\cos\theta = \frac{4B_L + B_T}{4B_L - B_T} \tag{4.48}$$

We see immediately that for real values of θ ($\cos\theta \leq 1$), B_T and B_L must have opposite signs, making the denominator greater than the numerator.

Figure 4.16 shows one possible case, in which $\omega_T < \omega_L$. In this case the pass band (θ real) starts at the lower cutoff frequency $\omega = \omega_T$ at which B_T is zero, $\cos\theta = 1$ (from (4.48)) and $\theta = 0$, and extends up to the upper cutoff frequency $\omega = \omega_L$ at which $B_L = 0$, $\cos\theta = -1$ and $\theta = \pi$.

* In case the filter has a large fractional bandwidth, it may be worth while to use the accurate lumped-circuit forms

$$B_T = \omega_T C_T(\omega/\omega_T - \omega_T/\omega) \tag{4.46a}$$

$$B_L = \omega_L C_L(\omega/\omega_L - \omega_L/\omega) \tag{4.46b}$$

The shape of the phase curves will depend on the relative rates of variation of B_T and B_L with frequency. Assuming the linear variations with frequency of (4.46) and (4.47) the shapes can be computed. This has been done for $C_L/C_T = 1, 3, 10$ and the results are shown in Fig. 4.17.

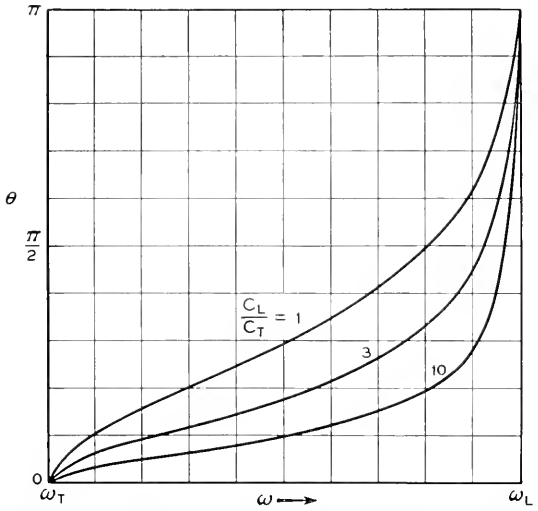

Fig. 4.17—Phase shift per section, θ, vs radian frequency ω for the conditions of Fig. 4.16.

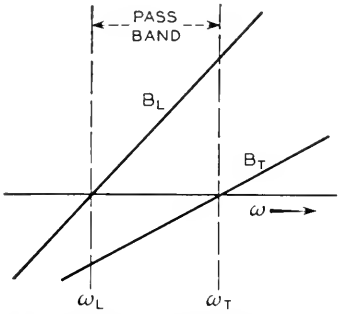

Fig. 4.18—Longitudinal and transverse susceptances which give $-\pi$ radians phase shift at the lower cutoff ($\omega = \omega_L$) and 0 degrees phase shift at the upper cutoff ($\omega = \omega_T$). This means a negative phase velocity.

It is of course possible to make $\omega_L > \omega_T$. In this case the situation is as shown in Fig. 4.18, the pass band extending from ω_L to ω_T. At $\omega = \omega_L$, $\cos\theta = -1$, $\theta = -\pi$. At $\omega = \omega_T$, $\cos\theta = 1$ and $\theta = 0$. In Fig. 4.19, assuming (4.46) and (4.47), θ has been plotted vs ω for $C_L/C_T = 1, 3, 10$.

The curves of Figs. 4.17 and 4.18 are not exact for any physical structure of the type shown in Fig. 4.10. In lumped circuit terms, they neglect coupling

between slots. They will be most accurate for structures with slots longitudinally far apart compared with the transverse dimensions, and least accurate for structures with slots close together. They do, however, form a valuable guide in understanding the performance of such structures and in evaluating the effect of the ratio of energies stored in the fields at the two cutoff frequencies.

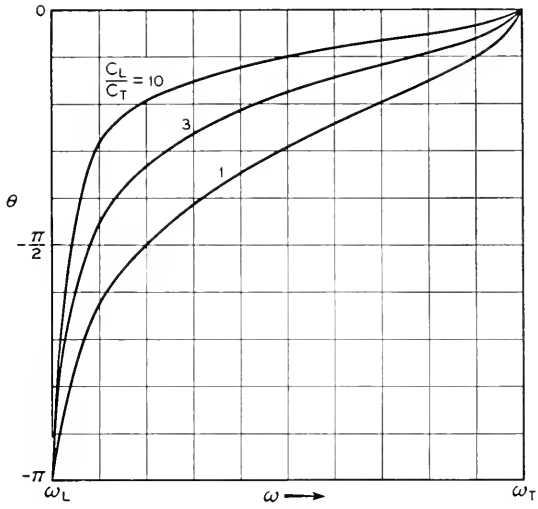

Fig. 4.19—Phase shift per section, θ, vs radian frequency, ω, for the conditions of Fig 4.18.

It is most likely that the voltages across the slots would be of most interest in connection with the circuit shown in Fig. 4.10. We can rewrite (4.41) in terms of B_T and B_L

$$V^2/P = \frac{1}{2(1 - 4B_L/B_T)(-B_T B_L)^{1/2}} \tag{4.49}$$

We see that V^2/P goes to 0 at $B_T = 0$ ($\omega = \omega_T$) and to infinity at $B_L = 0$ ($\omega = \omega_L$). In Fig. 4.20 assuming (4.46) and (4.47), $(V^2/P)(\omega_L C_L \omega_T C_T)$ is plotted vs ω for $C_L/C_T = 1, 3, 10$.

Let us consider another circuit, that shown in Fig. 4.11. We see that this consists of a number of resonators coupled together inductively. We might draw the equivalent circuits of these resonators as shown in Fig. 4.21. Here L and C are the effective inductance and the effective capacitance of the resonators without irises. They are chosen so that the resonant frequency ω_0 is given by

$$\omega_0 = \sqrt{LC} \tag{4.50}$$

and the variation of gap susceptance B with frequency is

$$\partial B/\partial\omega = 2C \tag{4.51}$$

The arrows show directions of current flow when the currents in the gap capacitances are all the same.

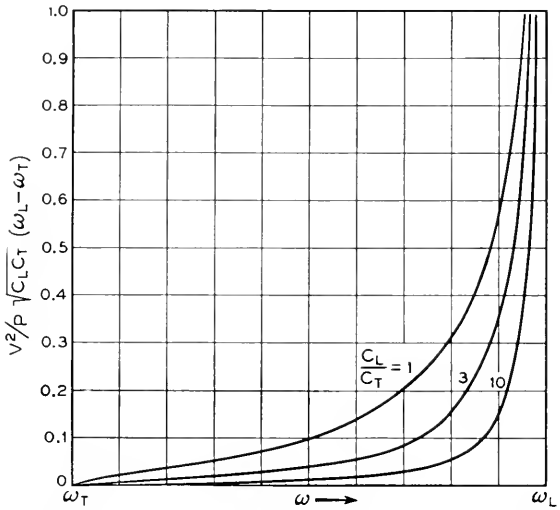

Fig. 4.20—A quantity proportional to $(E^2/\beta^2 P)$ vs ω for the conditions of Figs. 4.16 and 4.17.

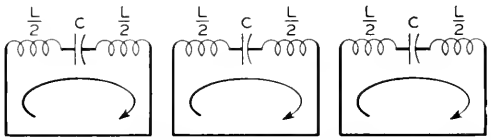

Fig. 4.21—A representation of the resonators of Fig. 4.11.

We can now represent the circuit of Fig. 4.11 by interconnecting the circuits of Fig. 4.21 by means of inductances L_M of Fig. 4.22. This gives a suitable representation, but one which is open to a minor objection: the gap capacitance does not appear across either a shunt or a series arm.

It is important to notice that there is another equally good representation, and there are probably many more. Suppose we draw the resonators as shown in Fig. 4.23 instead of as in Fig. 4.21. The inductance L and capacitance C are still properly given by 4.50 and 4.51. We can now interconnect the resonators inductively as shown in Fig. 4.24.

We should note one thing. In Fig. 4.21, the currents which are to flow in the common inductances of Fig. 4.22 flow in opposite directions when the

gap currents are in the same directions. In the representation of Fig. 4.23 the currents which will flow in the common inductances of Fig. 4.24 have been drawn in opposite directions, and we see that the currents in the gap capacitances flow alternately up and down. In other words, in Fig. 4.24, every other gap appears inverted. This can be taken into account by adding a phase angle $-\pi$ to θ as computed from (4.48).

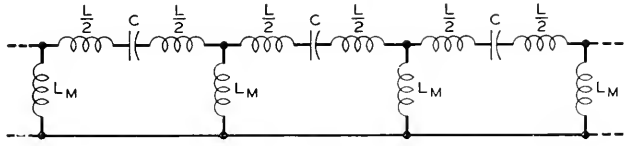

Fig. 4.22—The resonators of Fig. 4.11 coupled inductively.

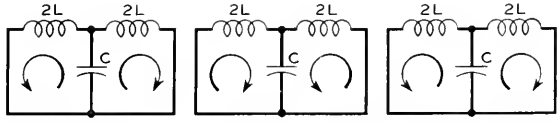

Fig. 4.23—Another representation of the resonators of Fig. 4.11.

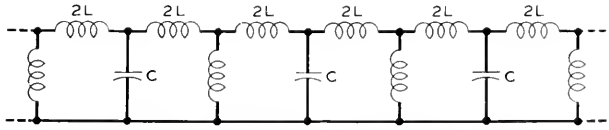

Fig. 4.24—Figure 4.23 with inductive coupling added.

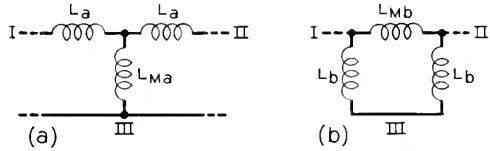

Fig. 4.25—A $T - \pi$ transformation used in connection with the circuit of Fig. 4.24.

Now, the T configuration of inductances in a of Fig. 4.25 can be replaced by the π configuration, b of Fig. 4.25. Imagine I and II to be connected together and a voltage to be applied between them and III. We see that

$$L_b = L_a + 2L_{Ma} \tag{4.52}$$

Imagine a voltage to be applied between I and II. We see that

$$1/L_a = 1/L_b + 2/L_{Mb} \tag{4.53}$$

If $L_{Ma} \ll L_a$, then L_b will be nearly equal to L_a and $L_{Mb} \gg L_b$.

By means of such a $T - \pi$ transformation we can redraw the equivalent circuit of Fig. 4.24 as shown in Fig. 4.26. The series susceptance B_1 is now

that of L_1, and the shunt susceptance is now that of the shunt resonant circuit consisting of C_2 (the effective capacitance of the resonators) and L_2.

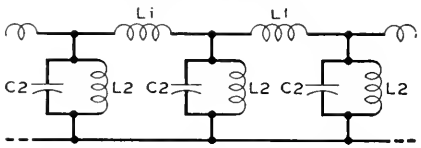

Fig. 4.26—The final representation of the circuit of Fig. 4.11.

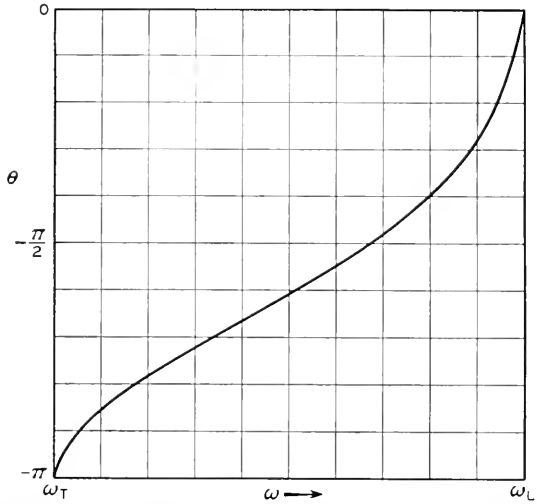

Fig. 4.27—The phase characteristic of the circuit of Fig. 4.11.

The transverse resonance, $B_2 = 0$, occurs at a frequency

$$\omega_T = \sqrt{C_2 L_2} \tag{4.54}$$

Near this frequency the transverse susceptance is given by

$$B_T = 2C_2(\omega - \omega_T) \tag{4.55}$$

The longitudinal resonance occurs at a frequency

$$\omega_L = \sqrt{2C_2 L_1 L_2/(L_1 + 2L_2)} \tag{4.56}$$

and near ω_L,

$$B_L = C_2(\omega - \omega_L) \tag{4.57}$$

These are just the forms we found in connection with the structure of Fig. 4.10; but we see that, in the case of the circuit of Fig. 4.11, the effective transverse capacitance is always twice the effective longitudinal capacitance ($C_L/C_T = 1/2$ in Fig. 4.19), and that $\omega_L > \omega_T$ for attainable volume of L_1.

We obtain θ vs ω by adding $-\pi$ to the phase angle from 4.48, using (4.55) and (4.57) in obtaining B_T and B_L. The phase angle vs. frequency is shown in Fig. 4.27. As the irises are made larger, the bandwidth, $\omega_L - \omega_T$, becomes larger, largely by a decrease in ω_L.

The voltage of interest is that across C_2, that is, that across the gap. From (4.37), (4.44), (4.45), (4.55) and (4.57) we obtain

$$V^2/P = 2/(-B_T B_L)^{1/2} \tag{4.58}$$

$$V^2/P = (\sqrt{2}/C_2)((\omega_L - \omega)(\omega - \omega_T))^{-1/2} \tag{4.59}$$

This goes to infinity at both $\omega = \omega_L$ and $\omega = \omega_T$. In Fig. 4.28, $(V^2/P)C_2\sqrt{\omega_L \omega_T}$ is plotted vs ω. This curve represents the performance of all narrow band structures of the type shown in Fig. 4.11.

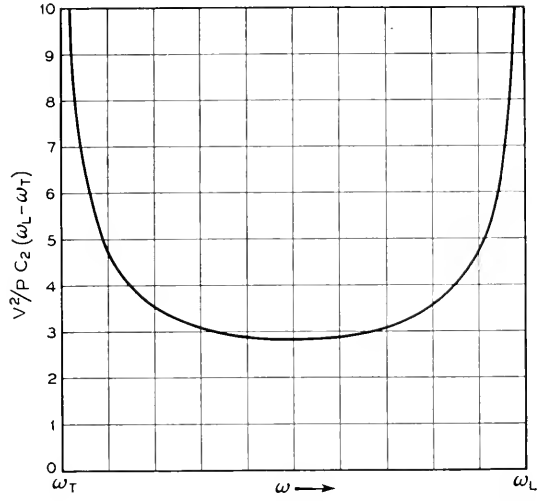

Fig. 4.28—A quantity proportional to $(E^2/\beta^2 P)$ for the circuit of Fig. 4.11, plotted vs radian frequency ω.

In a structure such as that shown in Fig. 4.11, there is little coupling between sections which are not adjacent, and hence the lumped-circuit representation used is probably quite accurate, and is certainly more accurate than in structures such as that shown in Fig. 4.10.

Other structures could be analyzed, but it is believed that the examples given above adequately illustrate the general procedures which can be employed.

4.4 Traveling Field Components

Filter-type circuits produce fields which are certainly not sinusoidal with distance. Indeed, with a structure such as that shown in Fig. 4.11, the elec-

trons are acted upon only when they are very near to the gaps. It is possible to analyze the performance of traveling-wave tubes on this basis[5]. The chief conclusion of such an analysis is that highly accurate results can be obtained by expressing the field as a sum of traveling waves and taking into account only the wave which has a phase velocity near to the electron velocity. Of course this is satisfactory only if the velocities of the other components are quite different from the electron velocity (that is, different by a fraction several times the gain parameter C).

As an example, consider a traveling-wave tube in which the electron stream passes through tubular sections of radius a, as shown in Fig. 4.29, and is acted upon by voltages appearing across gaps of length ℓ spaced L apart.

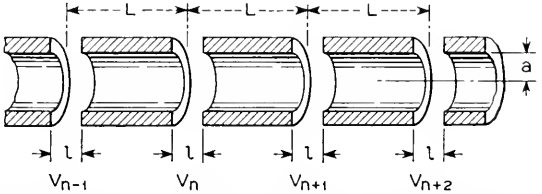

Fig. 4.29—A series of gaps in a tube of inside radius a. The gaps are ℓ long and are spaced L apart. Voltages V_n, etc., act across them.

A wave travels in some sort of structure and produces voltages across the gaps such that that across the nth gap, V, is

$$V_n = V_0 e^{-jn\theta} \tag{4.60}$$

where n is any integer.

We analyze this field into traveling-wave components which vary with distance as $\exp(-j\beta_m z)$ where

$$\beta_m = (\theta + 2m\pi)/L \tag{4.61}$$

where m is any positive or negative integer. Thus, the total field will be

$$E = \sum_{m=-\infty}^{\infty} E_m = \sum_{m=-\infty}^{\infty} A_m e^{-j\beta_m z} I_0(\gamma_m r) \tag{4.62}$$

$$\gamma_m^2 = \beta_m^2 - \beta_0^2 \tag{4.63}$$

Here $I_0(\gamma_m r)$ is a modified Bessel function, and γ_m has been chosen so that (4.62) satisfies Maxwell's equations.

[5] J. R. Pierce and Nelson Wax, "A Note on Filter-Type Traveling-Wave Amplifiers," *Proc. I.R.E.*, Vol. 37, pp. 622-625, June, 1949.

We will evaluate the coefficients by the usual means of Fourier analysis. Suppose we let $z = 0$ at the center of one of the gaps. We see that

$$\begin{aligned}\int_{-L/2}^{L/2} EE^* \, dz &= \sum_{m=-\infty}^{\infty} \int_{-L/2}^{L/2} A_m A_m^* I_0^2(\gamma_m r) \, dz \\ &= \sum_{m=-\infty}^{\infty} A_m A_m^* I_0^2(\gamma_m r) L\end{aligned} \tag{4.64}$$

All of the terms of the form $E_m E_p$, $p \neq m$ integrate to zero because the integral contains a term $\exp(-j2\pi(p - m)/L)z$.

Let us consider the field at the radius r. This is zero along the surface of the tube. We will assume with fair accuracy that it is constant and has a value $-V/\ell$ across the gap. Thus we have also at $r = a$,

$$\begin{aligned}\int_{-L/2}^{L/2} EE^* \, dz &= -(V/\ell) \sum_{m=-\infty}^{\infty} \int_{-\ell/2}^{\ell/2} A_m^* e^{-j\beta_m z} I_0(\gamma_m a) \, dz \\ &= -(V/\ell) \sum_{m=-\infty}^{\infty} (A_m^*) I_0(\gamma_m a) \left(\frac{e^{-j\beta_m \ell/2} - e^{j\beta_m \ell/2}}{j\beta} \right)\end{aligned} \tag{4.65}$$

We can rewrite this

$$\int_{-L/2}^{L/2} EE^* \, dz = -(V/\ell) \sum_{m=-\infty}^{\infty} A_m^* I_0(\gamma_m a) \frac{\sin(\beta_m \ell/2)}{(\beta_m \ell/2)} \tag{4.66}$$

By comparison with (4.64) we see that

$$A_m = -(V/L)(\sin(\beta_m \ell/2)/(\beta_m \ell/2))(1/I_0(\gamma a)) \tag{4.67}$$

This is the magnitude of the mth field component on the axis. The magnitude of the field at a radius r would be $I_0(\gamma r)$ times this.

The quantity $\beta_m \ell$ is an angle which we will call θ_g, the gap angle. Usually we are concerned with only a single field component, and hence can merely write γ instead of γ_m. Thus, we say that the magnitude E of the travelling field produced by a voltage V acting at intervals L is

$$E = -M(V/L) \tag{4.68}$$

$$M = \frac{\sin(\theta g/2)}{(\theta g/2)} \frac{I_0(\gamma r)}{I_0(\gamma a)} \tag{4.69}$$

$$\theta g = \beta \ell \tag{4.70}$$

The factor M is called the gap factor or the modulation coefficient*. For slow waves, γ is very nearly equal to β, and we can replace γr and γa by βr and βa. For unattenuated waves, M is a real positive number; and,

* This factor is often designated by β, but we have used β otherwise.

for the slowly varying waves with which we deal, we will always consider M as a real number.

The gap factor for some other physical arrangements is of interest. At a distance y above the two-dimensional array of strip electrodes shown in Fig. 4.30

$$M = \frac{\sin(\theta g/2)}{(\theta g/2)} e^{-\gamma y} \tag{4.71}$$

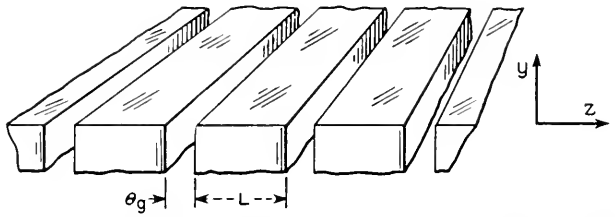

Fig. 4.30—A series of slots θ_g radians long separated by walls L long.

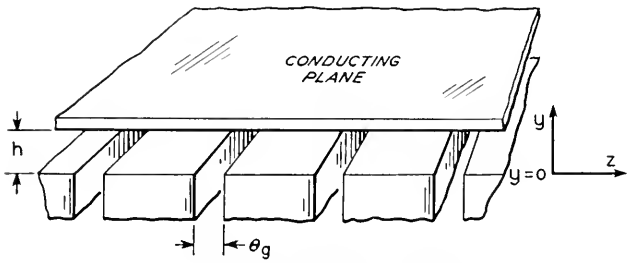

Fig. 4.31—A system similar to that of Fig. 4.30 but with the addition of an opposed conducting plane.

If we add a conducting plane a at $y = h$, as in Fig. 4.31,

$$M = \frac{\sin(\theta g/2)}{(\theta g/2)} \frac{\sinh \gamma(h - y)}{\sinh \gamma h} \tag{4.72}$$

For a symmetrical two-dimensional array, as shown in Fig. 4.32, with a separation of 2 h in the y direction and the fields above equal to the fields below

$$M = \frac{\sin(\theta g/2)}{(\theta g/2)} \frac{\cosh \gamma y}{\cosh \gamma h} \tag{4.73}$$

4.5 Effective Field and Effective Current

In Section 4.4 we have expressed a field component or "effective field" in terms of circuit voltage by means of a gap-factor or modulation coeffi-

cient M. This enables us to make calculations in terms of fields and currents at the electron stream.

The gap factor can be used in another way. A voltage appears across a gap, and the electron stream induces a current at the gap. At the electron stream the power P_1, produced in a distance L by a convection current i with the same z-variation as the field component considered, acting on the field component is

$$\begin{aligned} P_1 &= -Ei^*L \\ &= +(MV)i^* \end{aligned} \tag{4.74}$$

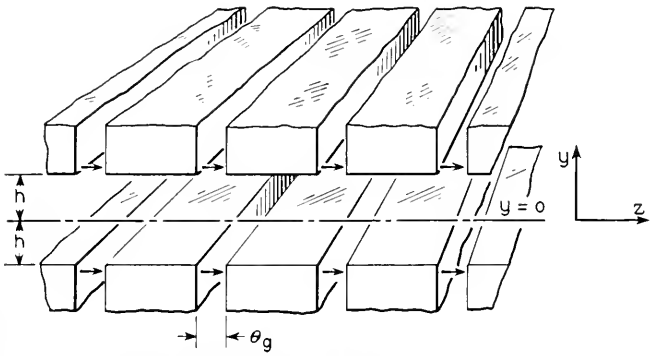

Fig. 4.32—A system of two opposed sets of slots.

At the circuit we observe some impressed current I flowing against the voltage V to produce a power

$$P_2 = VI^* \tag{4.75}$$

By the conservation of energy, these two powers must be the same, and we deduce that

$$I^* = Mi^* \tag{4.76}$$

or, since we take M as a real number

$$I = Mi \tag{4.77}$$

Thus, we have our choice of making calculations in terms of the beam current and a field component or effective field, or in terms of circuit voltage and an effective current, and in either case we make use of the modulation coefficient M.

Our gain parameter C^3 will be

$$C^3 = (V/L)^2M^2I_0/8\beta^2V_0$$

where V is circuit voltage. We can regard this in two ways. We can think of $-(V/L)M$ as the effective field at the location of the current I_0, or we can think of $M^2 I_0$ as the effective current referred to the circuit.

If we have a broad beam of electrons and a constant current density J_0 we compute (essentially as in Chapter III) a value of C^3 by integrating

$$C^3 = (1/8\beta^2 V_0) J_0 (V/L)^2 \int M^2 \, d\sigma \tag{4.78}$$

where $d\sigma$ is an element of area. We can think of the result in terms of an effective field E_e

$$E_e^2 = (V/L)^2 \frac{\int M^2 \, d\sigma}{\sigma} \tag{4.79}$$

where σ is the total beam area, and a total current σJ_0, or we can think of the integral (4.77) in terms of an effective current I_0 given by

$$I_0 = J_0 \int M^2 \, d\sigma \tag{4.80}$$

and the voltage at the circuit.

Of course, these same considerations apply to distributed circuits. Sometimes it is most convenient to think in terms of the total current and an effective field (as we did in connection with helices in Chapter III) and sometimes it is most convenient to think of the field at the circuit and an effective current. Either concept refers to the same mathematics.

4.6 Harmonic Operation

Of the field components making up E in (4.62) it is customary to regard the $m = 0$ component, for which $\beta = \theta/L$, as the *fundamental* field component, and the other components as *harmonic* components. These are sometimes called *Hartree harmonics*. If the electron speed is so adjusted that the interaction is with the $m = 0$ or fundamental component we have fundamental operation; if the electron speed is adjusted so that we have interaction with a harmonic component, we have harmonic operation.

There are several reasons for using harmonic operation in connection with filter-type circuits. For one thing the fundamental component may appear to be traveling backwards. Thus, for circuits of the type shown in Fig. 4.11, we see from Fig. 4.27 that θ is always negative. Now, in terms of the velocity v

$$\beta = \omega/v = \theta/L \tag{4.81}$$

and if θ is negative, v must be negative. However, consider the $m = 1$ component

$$\beta = \omega/v = (2\pi + \theta)/L \tag{4.82}$$

We see that, for this component, v is positive.

The interaction of electrons with backward-traveling field components will be considered later. Here it will merely be said that, in order to avoid interaction with waves traveling in both directions, one must avoid having the electron speed lie near both the speed of a forward component and the speed of a backward component.

In order that the fundamental component be slow, θ must be large or L must be small. The largest value of θ is that near one edge of the band, where θ approaches π. Thus, the largest fundamental value of β is π/L, and to make

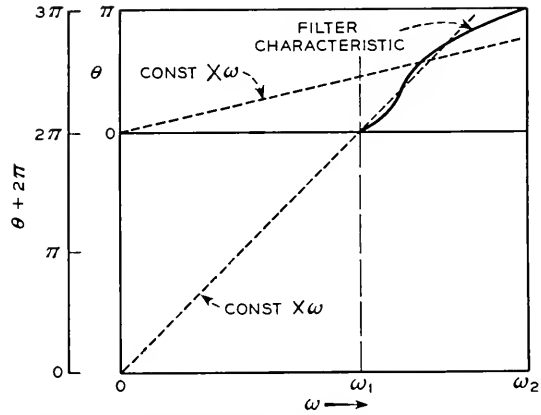

Fig. 4.33—The variation of phase with frequency for the fundamental (0 to π over the band) and a spatial harmonic (2π to 3π over the band). The dotted lines show ω divided by the electron velocity for the two cases. For amplification over a broad band the dotted curve should not depart much from the filter characteristic.

β large with $m = 0$ we must make L small and put the resonators very close together. This may be physically difficult or even impossible in tubes for very high frequencies. The alternative is to use a harmonic component, for which $\beta = (2m\pi + \theta)/L$.

Another reason for using harmonic operation is to achieve broad-band operation. The phase of a filter-type circuit changes by π radians between the lower cutoff frequency ω_1 and the upper cutoff frequency ω_2†. Now, for the wave velocity to be near to the electron velocity over a good part of the band, β must be nearly a constant times ω. Figure 4.33 shows how this can be approximately true for the $m = 1$ component even when it obviously won't be for the $m = 0$ or fundamental component. Similarly, for a filter with a narrower fractional bandwidth and hence a steeper curve of θ vs ω, a larger value of m might give a nearly constant value of v.

† The phase of some filters changes more than this, but they don't seem good candidates for traveling-wave tube circuits.

CHAPTER V

GENERAL CIRCUIT CONSIDERATIONS

Synopsis of Chapter

IN CHAPTERS III AND IV, helices and filter-type circuits have been considered. Other slow-wave circuits have been proposed, as, for instance, wave guides loaded continuously with dielectric material. One may ask what the best type of circuit is, or, indeed, in just what way do bad circuits differ from good circuits.

So far, we have as one criterion for a good circuit a high impedance, that is, a high value of $E^2/\beta^2 P$. If we want a broad-band amplifier we must have a constant phase velocity; that is, β must be proportional to frequency. Thus, two desirable circuit properties are: high impedance and constancy of phase velocity.

Now, $E^2/\beta^2 P$ can be written in the form

$$E^2/\beta^2 P = E^2/\beta^2 W v_g$$

where W is the stored energy per unit length for a field strength E, and v_g is the group velocity.

One way of making $E^2/\beta^2 P$ large is to make the stored energy for a given field strength small. In an electromagnetic wave, half of the stored energy is electric and half is magnetic. Thus, to make the total stored energy for a given field strength small we must make the energy stored in the electric field small. The energy stored in the electric field will be increased by the presence of material of a high dielectric constant, or by the presence of large opposed metallic surfaces, as in the circuits of Figs. 4.8 and 4.9. Thus, such circuits are poor as regards circuit impedance, however good they may be in other respects.

If the stored energy for a given field strength is held constant, $E^2/\beta^2 P$ may be increased by decreasing the group velocity. It is the phase velocity v which should match the electron speed. The group velocity v_g is given in terms of the phase velocity by (5.12). We see that the group velocity may be much smaller than the phase velocity if $-\partial v/\partial\omega$ is large. It is, for instance, a low group velocity near cutoff that accounts for the high impedance regions exhibited in Figs. 4.20 and 4.28. We remember, however, that, if the phase velocity of the circuit of a traveling-wave tube changes with frequency, the tube will have a narrow bandwidth, and thus the high

impedances attained through large values of $-\partial v/\partial\omega$ are useful over a narrow range of frequency only.

If we consider a broad electron stream of current density J_0, the highest effective value of $E^2/\beta^2 P$, and hence the highest value of C, will be attained if there is current everywhere that there is electric field, and if all of the electric field is longitudinal. This leads to a limiting value of C, which is given by (5.23). There λ_0 is the free-space wavelength. The nearest practical approach to this condition is perhaps a helix of fine wire flooded inside and outside with electrons.

In many cases, it is desirable to consider circuits for use with a narrow beam of electrons, over which the field may be taken as constant. As the helix is a common as well as a very good circuit, it might seem desirable to use it as a standard for comparison. However, the group velocity of the helix differs a little from the phase velocity, and it seems desirable instead to use a sort of hypothetical circuit or field for which the stored energy is almost the same as in the helix, but for which the group velocity is the same as the phase velocity. This has been referred to in the text as a "forced sinusoidal field." In Fig. 5.3, $(E^2/\beta^2 P)^{1/3}$ for the forced sinusoidal field is compared with $(E^2/\beta^2 P)^{1/3}$ for the helix.

Several other circuits are compared with this: the circular resonators of Fig. 5.4 (the square resonators of Fig. 5.4 give nearly the same impedance) and the resonant quarter-wave and half-wave wires of Figs. 5.6 and 5.7. The comparison is made in Fig. 5.8 for three voltages, which fix three phase velocities. In each case it is assumed that in some way the group velocity has been made equal to the phase velocity. Thus, the comparison is made on the basis of stored energies. The field is taken as the field at radius a (corresponding to the surface of the helix) in the case of the forced sinusoidal field, and at the point of highest field in the case of the resonators.

We see from Figs. 5.8 and 5.3 that a helix of small radius is a very fine circuit.

In circuits made up of a series of resonators, the group velocity can be changed within wide limits by varying the coupling between resonators, as by putting inductive or capacitive irises between them. Thus, even circuits with a large stored energy can be made to have a high impedance by sacrificing bandwidth.

The circuits of Fig. 5.4 have a large stored energy because of the large opposed surfaces. The wires of Fig. 5.6 have a small stored energy associated entirely with "fringing fields" about the wires. The narrow strips of Fig. 5.5 have about as much stored energy between the opposed flat surfaces as that in the fringing field, and are about as good as the half-wave wires of Fig. 5.7.

An actual circuit made up of resonators such as those of Fig. 5.4 will be

worse than Fig. 5.8 implies. Thus, there is a decrease of $(E^2/\beta^2 P)^{1/3}$ due to wall thickness. Thickening the flat opposed walls of the resonators decreases the spacing between the opposed surfaces, increases the capacitance and hence increases the stored energy for a given gap voltage. In Fig. 5.9 the factor f by which $(E^2/\beta^2 P)^{1/3}$ is reduced is plotted vs. the ratio of the wall thickness t to the resonator spacing L.

There is a further reduction of effective field because of the electrical length, θ in radians, of the space between opposed resonator surfaces. The lower curve in Fig. 5.10 gives a factor by which $(E^2/\beta^2 P)^{1/3}$ is reduced because of this. If the resonator spacing, θ_t in radians, is greater than 2.33 radians, it is best to make the opening, or space between the walls, only 2.33 radians long by making the opposed disks forming the walls very thick.

There is of course a further loss in effective field, both in the helix and in circuits made up of resonators, because of the falling-off of the field toward the center of the aperture through which the electrons pass. This was discussed in Chapter IV.

Finally, it should be pointed out that the fraction of the stored energy dissipated in losses during each cycle is inversely proportional to the Q of the circuit or of the resonators forming it. The distance the energy travels in a cycle is proportional to the group velocity. Thus, for a given Q the signal will decay more rapidly with distance if the group velocity is lowered (to increase $E^2/\beta^2 P$). Equations (5.38), (5.42) and (5.44) pertain to attenuation expressed in terms of group velocity. The table at the end of the chapter shows that a circuit made up of resonators and having a low enough group velocity to give it an impedance comparable with that of a helix can have a very high attenuation.

5.1 Group and Phase Velocity

Suppose we use a broad video pulse $F(t)$, containing radian frequencies p lying in the range 0 to p_0, to modulate a radio-frequency signal of radian frequency ω which is much larger than p_0, so as to give a radio-frequency pulse $f(t)$

$$f(t) = e^{j\omega t} F(t) \tag{5.1}$$

the functions $F(t)$ and $f(t)$ are indicated in Fig. 5.1.

$F(t)$, which is a real function of time, can be expressed by means of its Fourier transform in terms of its frequency components

$$F(t) = \int_{-p_0}^{p_0} A(p) e^{jpt}\, dp \tag{5.2}$$

Here $A(p)$ is a complex function of p, such that $A(-p)$ is the complex conjugate of $A(p)$ (this assures that $F(t)$ is real).

With $F(t)$ expressed as in (5.2), we can rewrite (5.1)

$$f(t) = \int_{-p_0}^{p_0} A(p) e^{j(\omega+p)t}\, dp \tag{5.3}$$

Now, suppose, as indicated in Fig. 5.2, we apply the *r-f* pulse $f(t)$ to the input of a transmission system of length L with a phase constant β which

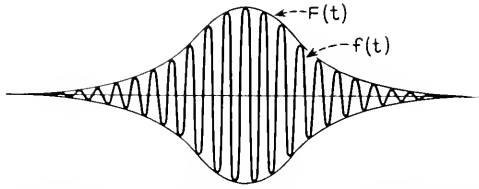

Fig. 5.1—A radio-frequency pulse varying with time as $f(t)$. The envelope varies with time as $F(t)$. The pulse might be produced by modulating a radio-frequency source with $F(t)$.

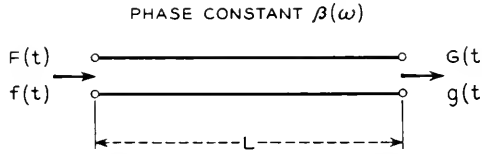

Fig. 5.2—When the pulse of Fig. 5.1 is applied to a transmission system of length L and phase constant $\beta(\omega)$ (a function of ω), the output pulse $g(t)$ has an envelope $G(t)$.

is a function of frequency. Let us assume that the system is lossless. The output $g(t)$ will then be

$$g(t) = \int_{-p_0}^{p_0} A(p) e^{j((\omega+p)t-\beta L)}\, dp \tag{5.4}$$

We have assumed that p_0 is much smaller than ω. Let us assume that over the range $\omega - p_0$ to $\omega + p_0$, β can be adequately represented by

$$\beta = \beta_0 + \frac{\partial \beta}{\partial \omega} p \tag{5.5}$$

In this case we obtain

$$g(t) = e^{j(\omega t-\beta_0 L)} \int_{-p_0}^{p_0} A(p)\, e^{jp(t-(\partial\beta/\partial\omega)L)}\, dp \tag{5.6}$$

The envelope at the output is

$$G(t) = \int_{-p_0}^{p_0} A(p)\, e^{jp(t-(\partial\beta/\partial\omega)L)}\, dp \tag{5.7}$$

By comparing this with (5.2) we see that

$$G(t) = F\left(t - \frac{\partial\beta}{\partial\omega} L\right) \tag{5.8}$$

In other words, the envelope at the output is of the same shape as at the input, but arrives a time τ later

$$\tau = \frac{\partial\beta}{\partial\omega} L \tag{5.9}$$

This implies that it travels with a velocity v_g

$$v_g = L/\tau = \left(\frac{\partial\beta}{\partial\omega}\right)^{-1} \tag{5.10}$$

This velocity is called the group velocity, because in a sense it is the velocity with which the group of frequency components making up the pulse travels down the circuit. It is certainly the velocity with which the energy stored in the electric and magnetic fields of the circuit travels; we could observe physically that, if at one time this energy is at a position x, a time t later it is at a position $x + v_g t$.

If the attenuation of the transmission circuit varies with frequency, the pulse shape will become distorted as the pulse travels and the group velocity loses its clear meaning. It is unlikely, however, that we shall go far wrong in using the concept of group velocity in connection with actual circuits.

We have used earlier the concept of phase velocity, which we have designated simply as v. In terms of phase velocity,

$$\beta = \frac{\omega}{v} \tag{5.11}$$

We see from (5.10) that in terms of phase velocity v the group velocity v_g is

$$v_g = v\left(1 - \frac{\omega}{v}\frac{\partial v}{\partial\omega}\right)^{-1} \tag{5.12}$$

For interaction of electrons with a wave to give gain in a traveling-wave tube, the electrons must have a velocity near the phase velocity v. Hence, for gain over a broad band of frequencies, v must not change with frequency; and if v does not change with frequency, then, from (5.12), $v_g = v$.

We note that the various harmonic components in a filter-type circuit have different phase velocities, some positive and some negative. The group

velocity is of course the same for all components, as they are all aspects of one wave. Relation (4.61) is consistent with this:

$$\beta_m = (\theta + 2m\pi)/L \tag{4.61}$$

$$1/v_g = \partial\beta_m/\partial\omega = (\partial\theta/\partial\omega)/L \tag{5.13}$$

5.2 Gain and Bandwidth in a Traveling-Wave Tube

We can rewrite the impedance parameter $E^2/\beta^2 P$ in terms of stored energy per unit length W for a field strength E, and a group velocity v_g. If W is the stored energy per unit length, the power flow P is

$$P = Wv_g \tag{5.14}$$

and, accordingly, we have

$$E^2/\beta^2 P = E^2/\beta^2 W v_g \tag{5.15}$$

And, for the gain parameter, we will have

$$C = (E^2/\beta^2 W v_g)^{1/3}(I_0/8V_0)^{1/3} \tag{5.16}$$

For example, we see from Fig. 4.20 that $E^2/\beta^2 P$ for the circuit of Fig. 4.10 goes to infinity at the upper cut-off. From Fig. 4.17 we see that $\partial\theta/\partial\omega$, and hence $1/v_g$, go to infinity at the upper cutoff, accounting for the infinite impedance. We see also that $\partial\theta/\partial\omega$ goes to infinity at the lower cutoff, but there the slot voltage and hence the longitudinal field also go to zero and hence $E^2/\beta^2 P$ does not go to infinity but to zero instead.

In the case of the circuit of Fig. 4.11, the gap voltage and hence the longitudinal field are finite for unit stored energy at both cutoffs. As $\partial\theta/\partial\omega$ is infinite at both cutoffs, W/P and hence $E^2/\beta^2 P$ go to infinity at both cutoffs, as shown in Fig. 4.28.

To get high gain in a traveling-wave tube at a given frequency and voltage (the phase velocity is specified by voltage) we see from (5.16) that we must have either a small stored energy per unit length for unit longitudinal field, or a small group velocity, v_g.

To have amplification over a broad band of frequencies we must have the phase velocity v substantially equal to the electron velocity over a broad band of frequencies. This means that for very broad-band operation, v must be substantially constant and hence in a broad-band tube the group velocity will be substantially the same as the phase velocity.

If the group velocity is made smaller, so that the gain is increased, the range of frequencies over which the phase velocity is near to the electron velocity is necessarily decreased. Thus, for a given phase velocity, as the group velocity is made less the gain increases but the bandwidth decreases.

Particular circuits can be compared on the basis of $(E^2/\beta^2 P)$ and band-

width. We have discussed the impedance and phase or velocity curves in Chapters III and IV. Field[1] has compared a coiled waveguide structure with a series of apertured disks of comparable dimensions. Both of these structures must have about the same stored energy for a given field strength. He found the coiled waveguide to have a low gain and broad bandwidth as compared with the apertured disks. We explain this by saying that the particular coiled waveguide he considered had a higher group velocity than did the apertured disk structure. Further, if the coiled waveguide could be altered in some way so as to have the same group velocity as the apertured disk structure it would necessarily have substantially the same gain and bandwidth.

In another instance, Mr. O. J. Zobel of these Laboratories evaluated the effect of broad-banding a filter-type circuit for a traveling-wave tube by m-derivation. He found the same gain for any combination of m and bandwidth which made $v = v_g(\partial v/\partial\omega = 0)$. We see this is just a particular instance of a general rule. The same thing holds for any type of broad-banding, as, by harmonic operation.

5.3 A Comparison of Circuits

The group velocity, the phase velocity and the ratio of the two are parameters which are often easily controlled, as, by varying the coupling between resonators in a filter composed of a series of resonators. Moreover, these parameters can often be controlled without much affecting the stored energy per unit length. For instance, in a series of resonators coupled by loops or irises, such as the circuit of Fig. 4.11, the stored energy is not much affected by the loops or irises unless these are very large, but the phase and group velocities are greatly changed by small changes in coupling.

Let us, then, think of circuits in terms of stored energy, and regard the phase and group velocities and their ratio as adjustable parameters. We find that, when we do this, there are not many essentially different configurations which promise to be of much use in traveling-wave tubes, and it is easy to make comparisons between extreme examples of these configurations.

5.3a Uniform Current Density throughout Field

Suppose we have a uniform current density J_0 wherever there is longitudinal electric field. We might approximate this case by flooding a helix of very fine wire with current inside and outside, or by passing current through a series of flat resonators whose walls were grids of fine wire.

[1] Lester M. Field, "Some Slow-Wave Structures for Traveling-Wave Tubes," *Proc. I.R.E.*, Vol. 37, pp. 34–40, January 1949.

In the latter case, if resonators had parallel walls of very fine mesh normal to the direction of electron motion there would be substantially no transverse electric field. All the electric field representing stored energy would act on the electron stream. In this case, we would have

$$W = \frac{\epsilon}{2} \int E^2 \, d\Sigma \tag{5.17}$$

Here $d\Sigma$ is an elementary area normal to the direction of propagation. W given by this expression is the total electric and magnetic stored energy per unit length. Where E is less than its peak value, the magnetic energy makes up the difference.

In evaluating $E^2 I_0$ in (5.16) we will have as an effective value

$$(EI_0)_{\text{eff}} = J_0 \int E \, d\Sigma \tag{5.18}$$

Hence, we will have for the gain parameter C

$$C = \left(\frac{J_0 \int E^2 \, d\Sigma}{\left(\frac{\omega}{v}\right)^2 \left(\frac{\epsilon}{2} \int E^2 \, d\Sigma\right) v_0 (8V_0)} \right)^{1/3}$$

$$C = \left(\frac{J_0}{4 \left(\frac{\omega}{v}\right)^2 \epsilon v_0 V_0} \right)^{1/3} \tag{5.19}$$

It is of interest to put this in a slightly different form. Suppose λ_0 is the free-space wavelength. Then

$$\frac{\omega}{v} = \frac{2\pi}{\lambda_0} \frac{c}{v} \tag{5.20}$$

where c is the velocity of light

$$c = 3 \times 10^{10} \text{ cm/sec} = 3 \times 10^8 \text{ m/sec}$$

Further, we have for synchronism between the electron velocity u_0 and the phase velocity v

$$v^2 = 2\eta V_0 \tag{5.21}$$

Also

$$\begin{aligned} c &= 1/\sqrt{\mu\epsilon} \\ \epsilon &= 1/c\sqrt{\mu/\epsilon} \\ \sqrt{\mu/\epsilon} &= 377 \text{ ohms} \end{aligned} \tag{5.22}$$

Using (5.20), (5.21), (5.22) in connection with (5.19), we obtain

$$C = \left(\frac{\eta\sqrt{\mu/\epsilon}\ J_0\lambda_0^2}{16\pi^2 c v_g}\right)^{1/3} = 11.16\ (J_0\lambda_0^2/v_g)^{1/3} \tag{5.23}$$

We have in (5.23) an expression for the gain parameter C in case longitudinal fields only are present and in case there is a uniform current density J_0 wherever there is a longitudinal field.

In a number of cases, as in case of a large-diameter helix, or of a resonator with large apertures, the stored energy due to the transverse field is about equal to that due to the longitudinal field and C will be $2^{-1/3}$ times as great as the value of C given by (5.23). Thus, the value of C given by (5.23), or even $2^{-1/3}$ times this, represents an unattainable ideal. It is nevertheless of interest in indicating how limiting behavior depends on various parameters. For instance, we see that if the wavelength λ_0 is made shorter, a higher current density must be used if C is not to be lowered; for a constant C the current density must be such as to give a constant current through a square a wavelength on a side.

In the table below, some values of C have been computed from (5.23) for various wavelengths and current densities. The broad-band condition of equal phase and group velocities has been assumed, and the voltage has been taken as 1,000 volts.

Wavelength Cm \ Amp/cm²	.1	1
5	.060	.130
.5	.013	.028

For larger voltages, C will be smaller. C can of course be made larger by making the group velocity smaller than the phase velocity.

Of course, if the electron stream does not pass through some portions of the field, C will be smaller than given by (5.23). C will also be less if there are "harmonic" field components which do not vary in the z direction as $\exp(j\omega z/v)$.

5.3b Narrow Beams

Usually, no attempt is made to fill the entire field with electron flow even though this is necessary in getting a large value of C for a given current density. Instead a narrow electron beam is shot through a region of high

field. We then wish to relate the peak field strength to the stored energy in comparing various circuits.

Let us first consider a helically conducting sheet of radius a. The upper curve of Fig. 5.3 shows $(E^2/\beta^2 P)^{1/3}(v/c)^{1/3}$ vs. βa. In obtaining this curve it was assumed that $v \ll c$, so that γ can be taken as equal to β. The field E is the longitudinal field at the surface of the helically conducting cylinder. Figure 5.3 can be obtained from Fig. 3.4 by multiplying $F(\gamma a)$ by $(I_0(\gamma a))^{2/3}$ to give a curve valid for the field at $r = a$.

The helix has a very small circumferential electric field which represents "useless" stored energy. The lower curve of Fig. 5.3 is based on the stored electric energy of an axially symmetrical sinusoidal field impressed at the radius a.† This field has no circumferential component but is otherwise the

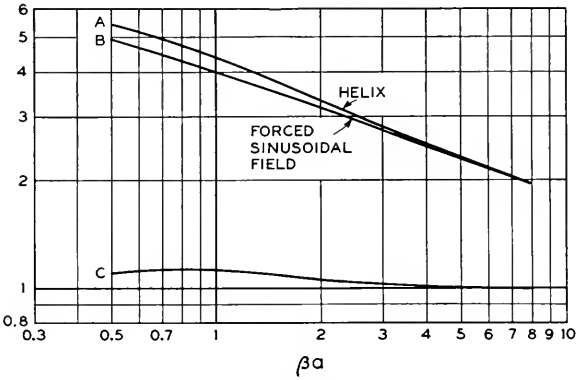

Fig. 5.3—The impedance parameter $(E^2/\beta^2 P)^{1/3}$ compared for a helically conducting sheet (A) and a forced sinusoidal field (B) with a group velocity equal to the phase velocity. The helix has a higher impedance because the phase velocity is higher than the group velocity by a radio shown to the $\frac{1}{3}$ power by curve C.

same as the electric field of the helix (again assuming $v \ll c$). We can imagine such a field propagating because of an inductive sheet at the radius a, which provides stored magnetic energy enough to make the electric and magnetic energies equal. The quantity plotted vs. βa is $(E^2/\beta^2 P)^{1/3}\ (v/c)^{1/3}\ (v_g/v)^{1/3}$.

The forced sinusoidal field is not the field of some particular circuit for which a certain group velocity v_g corresponds to a given phase velocity v. Hence, the factor $(v_g/v)^{1/3}$ is included in the ordinate, so that the curve will be the same no matter what group velocity is assumed. For the helically conducting sheet, a definite group velocity goes with a given phase velocity. In Fig. 5.3, the ordinate of the curve for the helically conducting sheet does not contain the factor $(v_g/v)^{1/3}$. If, for instance, we assume $v_g = v$

† See Appendix III.

in connection with the curve for the forced sinusoidal field, then the two ordinates are both $(E^2/\beta^2 P)^{1/3}\ (v/c)^{1/3}$ and the curve for the sheet is higher than that for the forced field because, for the helically conducting sheet,

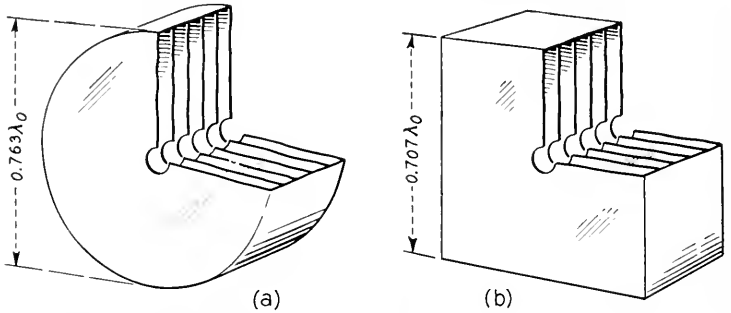

Fig. 5.4—Pillbox and rectangular resonators. When a number of resonators are coupled one to the next, a filter-type circuit is formed.

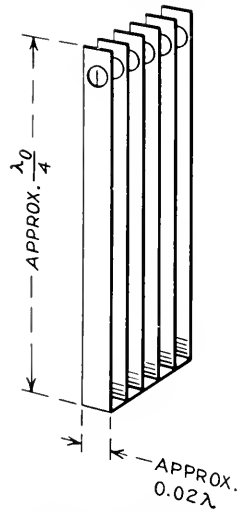

Fig. 5.5—Resonators with the opposing parallel surfaces reduced to lower stored energy and increase impedance.

$v_g < v$ for small values of γa. Curve C shows $(v/v_g)^{1/3}$ for the sheet vs. βa. Aside from the influence of group velocity, we might have expected the curve for the sheet to be a little lower than that for the forced field because of the energy associated with the transverse electric field component of the sheet. This, however, becomes small in comparison with the transverse magnetic component when $v \ll c$, as we have assumed.

Various other circuits will be compared, using the impressed sinusoidal field as a sort of standard of reference.

One of the circuits which will be considered is a series of flat resonators coupled together to make a filter. Figure 5.4a shows a series of very thin pillboxes with walls of negligible thickness. A small central hole is provided for the electron stream, and the field E is to be measured at the edge of this hole. The diameter is chosen to obtain resonance at a wavelength λ_0. Figure 5.4b shows a similar series of flat square resonators.

For the round resonators it is found that*

$$(E^2/\beta^2 P)^{1/3} = 5.36\ (v/c)^{1/3}\ (v/v_g)^{1/3} \tag{5.24}$$

for the square resonators*

$$(E^2/\beta^2 P)^{1/3} = 5.33\ (v/c)^{1/3}\ (v/v_g)^{1/3} \tag{5.25}$$

For practical purposes these are negligibly different.

* See Appendix III.

Suppose we wanted to improve on such circuits by reducing the stored energy. An obvious procedure would be to cut away most of the flat opposed surfaces as shown in Fig. 5.5. This reduces the energy stored between the resonator walls, but results in energy storage outside of the open edges, energy associated with a "fringing field."

Going to an extreme, we might consider an array of closely spaced very fine wires, as shown in Fig. 5.6. Here there are no opposed flat surfaces, and all of the electric field is a fringing field; we have reached an irreducible minimum of stored energy in paring down the resonator.

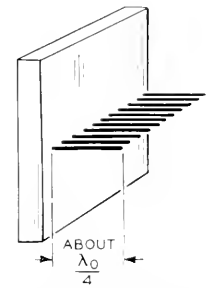

Fig. 5.6—Quarter-wave wires, which have a minimum of stored energy.

The structure of Fig. 5.6 has not been analyzed exactly, but that of Fig. 5.7 has. In Fig. 5.7, we have an array of fine, closely spaced half-wave wires between parallel planes.* This should have roughly twice the stored energy of Fig. 5.6, and we will estimate $(E^2/\beta^2 P)^{1/3}$ for Fig. 5.6 on this basis. We obtain in Appendix III:

For the half-wave wires,

$$(E^2/\beta^2 P)^{1/3} = 6.20\ (v/v_g)^{1/3} \qquad (5.25)$$

and hence for the quarter-wave wires, approximately

$$(E^2/\beta^2 P)^{1/3} = 7.81\ (v/v_g)^{1/3} \qquad (5.26)$$

As we have noted, $(v\ c)$, which appears in the expression for $(E^2/\beta^2 P)^{1/3}$ for the sinusoidal field impressed at radius a and in (5.24) and (5.25), is a

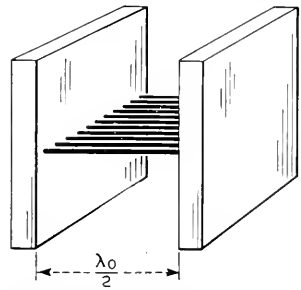

Fig. 5.7—Half-wave wires between parallel planes. The stored energy can be calculated for this configuration, assuming the wires to be very fine. The circuit does not propagate a wave unless added coupling is provided.

function of the accelerating voltage. Figure 5.8 makes a comparison between the sinusoidal field impressed at a radius a, curve A; the flat resonators, either circular or square, B; the half-wave wires, C; and the quarter-

* There is no transverse magnetic wave propagation along such a circuit unless extra coupling or loading is provided. Behavior of nonpropagating circuits in the presence of an electron stream is considered in Section 4 of Chapter XIV.

wave wires C'. In all cases, it is assumed that the coupling is so adjusted as to make $(v_g \cdot v) = 1$ (broad-band condition).

What sort of information can we get from the curves of Fig. 5.8? Consider the curves for 1,000 volts. Suppose we want to cut down the opposed areas of resonators, as indicated in Fig. 5.5, so as to make them as good as half-wave wires (curve C). The edge capacitance in Fig. 5.5 will be about equal to that for quarter-wave wires (curve C'). Curve C' is about 3.7 times as high as curve B, and hence represents only about $(1/3.7)^3 = .02$ as much capacitance. If we make the opposed area in Fig. 5.5 about .01 that in Fig. 5.4a or b, the capacitance* between opposed surfaces will equal the edge

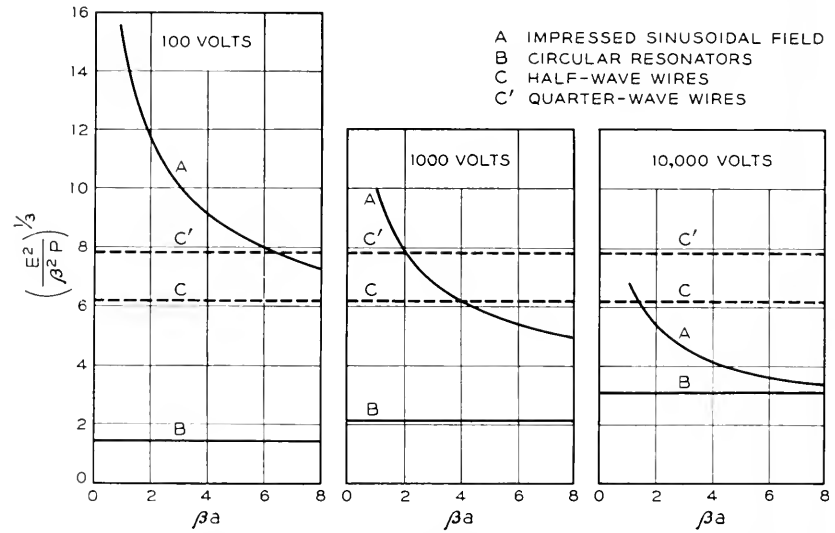

Fig. 5.8—Comparisons in terms of impedance parameter of an impressed sinusoidal field (A), circular resonators (B), half-wave wires (C) and quarter-wave wires (C') assuming the group and phase velocities to equal the electron velocity. The radius of the impressed sinusoidal field is a.

capacitance and the total stored energy will be twice that for quarter-wave wires, or equal to that for half-wave wires. This area is shown approximately to scale relative to Fig. 5.4 in Fig. 5.5. Thus, at 1,000 volts the resonant strips of Fig. 5.5 are about as good as fine, closely spaced half-wave wires.

Suppose again that we wish at 1,000 volts to make the gain of the resonators of Fig. 5.4 (or of a coiled waveguide) as good as that for a helix with $\beta a = 3$. For $\beta a = 3$ the helix curve A is about 3.2 times as high as the resona-

* This takes into account a difference in field distribution—that in Fig. 5.4b.

tor curve B. As $(E^2/\beta^2 P)^{1/3}$ varies as $(v/v_0)^{1/3}$, we must adjust the coupling between resonators so as to make

$$v_0 = v/(3.2)^3 = .031\ v$$

in order to make $(E^2/\beta^2 P)^{1/3}$ the same for the resonators as for the helix. From (5.12) we see that this means that a change in frequency by a fraction .002 must change v by a fraction .06. Ordinarily, a fractional variation of v of $\pm .03$ would cause a very serious falling off in gain. At 3,000 mc the total frequency variation of .002 times in v would be 6 mc. This is then a measure of the bandwidth of a series of resonators used in place of a helix for which $\beta a = 3$ and adjusted to give the same gain.

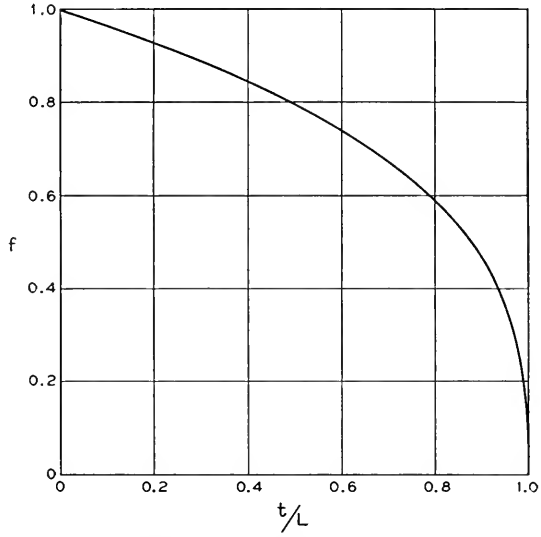

Fig. 5.9—The factor f by which $(E^2/\beta^2 P)^{1/3}$ for a series of resonators such as those of Fig. 5.4 is reduced because of wall thickness t, in relation to gap spacing L.

5.4 Physical Limitations

In Section 3.3b the resonators were assumed to be very thin and to have walls of zero thickness. Of course the walls must have finite thickness, and it is impractical to make the resonators extremely thin. The wall thickness and the finite transit time across the resonators both reduce $E^2/\beta^2 P$.

5.4a Effect of Wall Thickness

Consider the resonators of Fig. 5.4. Let L be the spacing between resonators ($1/L$ resonators per unit length), and t be the wall thickness. Thus, the gap length is $(L - t)$. Suppose we keep L and the voltage across each

resonator constant, so as to keep the field constant, but vary t. The capacitance will be proportional to $(L - t)^{-1}$ and, as the stored energy is the voltage squared times the capacitance, we see that $(E^2/\beta^2 P)^{1/3}$ will be reduced by a factor f,

$$f = (1 - t/L)^{1/3} \tag{5.27}$$

The factor f is plotted vs. t/L in Fig. 5.9.

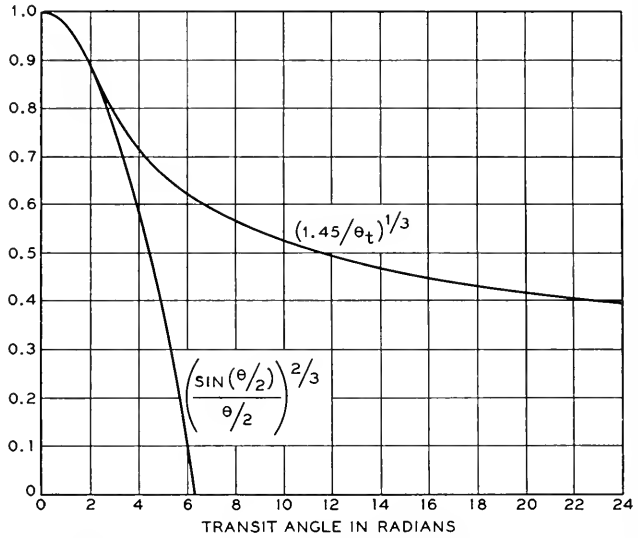

Fig. 5.10—The lower curve shows the factor by which $E^2/\beta^2 P$ is reduced by gap length, θ in radians. If the gap spacing is greater than 2.33 radians, it is best to make the gap 2.33 radians long. Then the upper curve applies.

5.4b Transit Time

As it is impractical to make the resonators infinitely thin, there will be some transit angle θ_g across the resonator, where

$$\theta_g = \beta \ell \tag{5.28}$$

Here ℓ is the space between resonator walls, or, the length of the gap. If we assume a uniform electric field between walls, the gap factor M, that is, the ratio of peak energy gained in electron volts to peak resonator voltage, or the ratio of the magnitude of the sinusoidal field component produced to that which would be produced by the same number of infinitely thin gaps with the same voltages, will be (from (4.69) with $r = a$)

$$M = \frac{\sin(\theta_g/2)}{\theta_g/2} \tag{5.29}$$

For a series of resonators θ_g long with infinitely thin walls $E^2/\beta^2 P$ will be less than the values given by (5.24) and (5.25) by a factor $M^{2/3}$. This is plotted vs. θ_g in Fig. 5.10.

5.4c Fixed Gap Spacing

Suppose it is decided in advance to put only one gap in a length specified by the transit angle θ_t. How wide should the gap be made, and how much will $E^2/\beta^2 P$ be reduced below the value for very thin resonators and infinitely thin walls?

Let us assume that all the stored energy is energy stored between parallel planes separated by the gap thickness, expressed in radians as θ or in distance as L

$$\theta_t = \beta \ell$$
$$\theta_g = \beta L$$

Here ℓ is the gap spacing and L is the spacing between resonators.

From Section 4.4 of Chapter IV we see that if V is the gap voltage, the field strength E is given by

$$E = MV/L$$

The stored energy per unit length, W, will be

$$W = W_0 V^2/\ell L \tag{5.30}$$

Here W_0 is a constant depending on the cross-section of the resonators. Thus, for unit field strength, the stored energy will be

$$\begin{aligned} W &= W_0 L/\ell M^2 \\ W &= W_0(\theta_t/\theta_g)(\theta_g/2)^2/\sin^2(\theta_g/2) \end{aligned} \tag{5.31}$$

We see that W_0 is merely the value of W when $\theta_t = \theta_g$ and $\theta_g = 0$, or, for zero wall thickness and very thin resonators. Thus, the ratio W/W_0 relates the actual stored energy per unit length per unit field to this optimum stored energy for resonators of the same cross section.

For $\theta_t < 2.33$, W/W_0 is smallest (best) for $\theta_g = \theta_t$ (zero wall thickness). For larger values of θ_t, the optimum value of θ_g is 2.33 radians and for this optimum value

$$(W_0/W)^{1/3} = (1.450/\theta_t)^{1/3} \tag{5.32}$$

If $\theta_t < 2.33$, it is thus best to make $\theta_g = \theta_t$. Then $(E^2/\beta^2 P)^{1/3}$ is reduced by the factor $[\sin(\theta/2)/(\theta/2)]^{2/3}$, which is plotted in Fig. 5.10. If $\theta_t > 2.33$, it is best to make $\theta = 2.33$. Then $(E^2/\beta^2 P)^{1/3}$ is reduced from the

value for thin resonators with infinitely thin walls by a factor given by (5.32), which is plotted vs. θ_t in Fig. 5.10.

If there are edge effects, the optimum gap spacing and the reduction in $(E^2/\beta^2 P)^{1/3}$ will be somewhat different. However, Fig. 5.10 should still be a useful guide.

In case of wide gap separation (large θ_t), there would be some gain in using reentrant resonators, as shown in Fig. 4.11, in order to reduce the capacitance. How good can such a structure be? Certainly, it will be worse than a helix. Consider merely the sections of metal tube with short gaps, which surround the electron beam. The shorter the gaps, the greater the capacitance. The space outside the beam has been capacitively loaded, which tends to reduce the impedance. This capacitance can be thought of as being associated with many spatial harmonics in the electric field, which do not contribute to interaction with the electrons.

5.5 Attenuation

Suppose we have a circuit made up of resonators with specified unloaded Q.† The energy lost per cycle is

$$W_L = 2\pi W_S/Q \tag{5.33}$$

In one cycle, however, a signal moves forward a distance L, where

$$L = v_g/f \tag{5.34}$$

The fractional energy loss per unit distance, which we will call 2α, is

$$2\alpha = \frac{W_L}{W_S}\frac{1}{L} \tag{5.35}$$

whence

$$\alpha = \frac{\omega}{2Qv_g} \tag{5.36}$$

So defined, α is the attenuation constant, and the amplitude will decay along the circuit as $\exp(-\alpha z)$.

The wavelength, λ, is given by

$$\lambda = v/f = 2\pi v/\omega \tag{5.37}$$

The loss per wavelength in db is

$$\begin{aligned} \text{db/wavelength} &= 20 \log_{10} \exp(\alpha\lambda) \\ \text{db/wavelength} &= \frac{27.3}{Q}\frac{v}{v_g} \end{aligned} \tag{5.38}$$

† Disregarding coupling losses, the circuit and the resonantors will both have this same Q.

We see that, for given values of v and Q, decreasing the group velocity, which increases $E^2/\beta^2 P$, also increases the attenuation per wavelength.

5.5a Attenuation of Circuits

For various structures, Q can be evaluated in terms of surface resistivity, R, the intrinsic resistance of space, $\sqrt{\mu/\epsilon} = 377$ ohms, and varous other parameters. For instance, Schelkunoff[2] gives for the Q of a pill-box resonator

$$Q = \frac{1.20(\sqrt{\mu/\epsilon}/R)}{1 + a/h} \tag{5.39}$$

Here a is the radius of the resonator and h is the height. If we express the radius in terms of the resonant wavelength λ_0 ($a = 1.2\lambda_0/\pi$), we obtain

$$Q = \frac{\pi(\sqrt{\mu/\epsilon}/R)(v/c)}{(1 + h/a)n} \tag{5.40}$$

Here n is the number of resonators per wavelength (assuming the walls separating the resonators to be of negligible thickness); thus

$$n = h/\lambda = (h/\lambda_0)(c/v) \tag{5.41}$$

From (5.40) and (5.38) we obtain for a series of pill-box resonators

$$\text{db/wavelength} = 8.68(R/\sqrt{\mu/\epsilon})(c/v_g)(1 + h/a)n \tag{5.42}$$

In Appendix III an estimate of the Q of an array of fine half-wave parallel wires is made by assuming conduction in one direction with a surface resistance R. On this basis, Q is found to be

$$Q = (\sqrt{\mu/\epsilon}/R)(v/c) \tag{5.43}$$

and hence

$$\text{db/wavelength} = 27.3(R/\sqrt{\mu/\epsilon})(c/v_g) \tag{5.44}$$

For non-magnetic materials, surface resistance varies as the square root of the resistivity times the frequency. The table below gives R for copper and db/wavelength for pill-box resonators for $h/a \ll 1$ (5.42) and for wires (5.44) for several frequencies

f, mc	R, Ohms	(db/wavelength)/ (c/v_g)	
		Pill-box Resonators	Wires
3,000	.0142	$3.3 \times 10^{-4}n$	10.3×10^{-4}
10,000	.0260	$6.0 \times 10^{-4}n$	18.1×10^{-4}
30,000	.0450	$10.4 \times 10^{-4}n$	32.6×10^{-4}

In Section 3.3b a circuit made up of resonators, with a group velocity .031 times the phase velocity, was discussed. Suppose such a circuit were

[2] Electromagnetic Waves, S. A. Schelkunoff, Van Nostrand, 1943. Page 269.

used at 1,000 volts ($c/v = 16.5$), were 40 wavelengths long, and had three copper resonators per wavelength. The total attenuation in db is given below

f, mc	Attenuation, db
3,000	21
10,000	38
30,000	67

CHAPTER VI

THE CIRCUIT DESCRIBED IN TERMS OF NORMAL MODES

Synopsis of Chapter

IN CHAPTER II, the field produced by the current in the electron stream, which was assumed to vary as $\exp(-\Gamma z)$, was deduced from a simple model in which the electron stream was assumed to be very close to an artificial line of susceptance B and reactance X per unit length. Following these assumptions, the voltage per unit length was found to be that of equation (2.10) and the field E in the z direction would accordingly be Γ times this, or

$$E = \frac{\Gamma^2 \Gamma_1 K}{\Gamma_1^2 - \Gamma^2} i \tag{6.1}$$

Here we will remember that Γ_1 is the natural propagation constant of the line, and K is the characteristic impedance.

We further replaced K by a quantity

$$E^2/\beta^2 P = 2K \tag{6.2}$$

where E is the field produced by a power flow P, and β is the phase constant of the line. For a lossless line, Γ_1 is a pure imaginary and

$$\beta^2 = -\Gamma_1^2 \tag{6.3}$$

From (6.1) and (6.2) we obtain

$$E = \frac{\Gamma^2 \Gamma_1 (E^2/\beta^2 P)}{2(\Gamma_1^2 - \Gamma^2)} i \tag{6.4}$$

To the writer it seems intuitively clear that the derivation of Chapter II is correct for waves with a phase velocity small compared with the velocity of light, and that (6.4) correctly gives the part of the field associated with the excitation of the circuit. However, it is clear that there are other field components excited; a bunched electron stream will produce a field even in the absence of a circuit. Further, many legitimate questions can be raised. For instance, in Chapter II capacitive coupling only was considered. What about mutual inductance between the electron stream and the inductances of the line?

The best procedure seems to be to analyze the situation in a way we know to be valid, and then to make such approximations as seem reasonable. One approximation we can make is, for instance, that the phase velocity of the wave is quite small compared with the speed of light, so that

$$| \Gamma_1 |^2 \gg \beta_0^2 = (\omega/c)^2 \tag{6.5}$$

In this chapter we shall consider a lossless circuit which supports a group of transverse magnetic modes of wave propagation. The finned structure of Fig. 4.3 is such a circuit, and so are the circuits of Figs. 4.8 and 4.9 (assuming that the fins are so closely spaced that the circuit can be regarded as smooth). It is assumed that waves are excited in such a circuit by a current in the z direction varying with distance as $\exp(-\Gamma z)$ and distributed normal to the z direction as a function of x and y, $\hat{J}(x, y)$. Such a current might arise from the bunching at low signal levels of a broad beam of electrons confined by a strong magnetic field so as not to move appreciably normal to the z direction.

The structure considered may support transverse electric waves, but these can be ignored because they will not be excited by the impressed current.

In the absence of an impressed current, any field distribution in the structure can be expressed as the sum of excitations of a number of pairs of normal modes of propagation. For one particular pair of modes, the field distribution normal to the z direction can be expressed in terms of a function $\hat{\pi}_n(x, y)$ and the field components will vary in the z direction as $\exp(\pm\Gamma_n z)$. Here the $+$ sign gives one mode of the pair and the $-$ sign the other. If Γ_n is real the mode is *passive*; the field decays exponentially with distance. If Γ_n is imaginary the mode is *active*; the field pattern of the mode propagates without loss in the z direction.

An impressed current which varies in the z direction as $\exp(-\Gamma z)$ will excite a field pattern which also varies in the z direction as $\exp(-\Gamma z)$, and as some function of x and y normal to the z direction. We may, if we wish, regard the variation of the field normal to the z direction as made up of a combination of the field patterns of the normal modes of propagation, the patterns specified by the functions $\hat{\pi}_n(x, y)$. Now, a pattern specified by $\hat{\pi}_n(x, y)$ coupled with a variation $\exp(\pm\Gamma_n z)$ in the z direction satisfies Maxwell's equations and the boundary conditions imposed by the circuit with *no* impressed current. If, however, we assume the same variation with x and y but a variation as $\exp(-\Gamma z)$ with z, Maxwell's equations will be satisfied only if there is an impressed current having a distribution normal to the z direction which also can be expressed by the function $\hat{\pi}_n(x, y)$.

Suppose we add up the various forced modes in such relative strength and phase that the total of the impressed currents associated with them is equal to the actual impressed current. Then, the sum of the fields of these

modes is the actual field produced by the actual impressed current. The field is so expressed in (6.44) where the current components J_n are defined by (6.36).

If it is assumed that there is only one mode of propagation, and if it is assumed that the field is constant over the electron flow, (6.44) can be put in the form shown in (6.47). For waves with a phase velocity small compared with the velocity of light, this reduces to (6.4), which was based on the simple circuit of Fig. 2.3.

Of course, actual circuits have, besides the one desired active mode, an infinity of passive modes and perhaps other active modes as well. In Chapter VII a way of taking these into account will be pointed out.

Actual circuits are certainly not lossless, and the fields of the helix, for instance, are not purely transverse magnetic fields. In such a case it is perhaps simplest to assume that the modes of propagation exist and to calculate the amount of excitation by energy transfer considerations. This has been done earlier[1], at first subject to the error of omitting a term which later[2] was added. In (6.55) of this chapter, (6.44) is reexpressed in a form suitable for comparison with this earlier work, and is found to agree.

Many circuits are not smooth in the z direction. The writer believes that usually small error will result from ignoring this fact, at least at low signal levels.

6.1 Excitation of Transverse Magnetic Modes of Propagation by a Longitudinal Current

We will consider here a system in which the natural modes of propagation are transverse magnetic waves. The circuit of Fig. 4.3, in which a slow wave is produced by finned structures, is an example. We will remember that the modes of propagation derived in Section 4.1 of Chapter IV were of this type. We will consider here that any structure the circuit may have (fins, for instance) is fine enough so that the circuit may be regarded as smooth in the z direction.

Any transverse electric modes which may exist in the structure will not be excited by longitudinal currents, and hence may be disregarded.

The analysis presented here will follow Chapter X of Schelkunoff's Electromagnetic Waves.

The divergence of the magnetic field H is zero. As there is no z component of field, we have

[1] J. R. Pierce, "Theory of the Beam-Type Traveling-Wave Tube," *Proc. I.R.E.*, Vol. 35, pp. 111–123, February, 1947.

[2] J. R. Pierce, "Effect of Passive Modes in Traveling-Wave Tubes," *Proc. I.R.E.*, Vol. 36, pp. 993–997, August, 1948.

$$\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} = 0 \tag{6.6}$$

This will be satisfied if we express the magnetic field in terms of a "stream function", π

$$H_x = \frac{\partial \pi}{\partial y} \tag{6.7}$$

$$H_y = -\frac{\partial \pi}{\partial x} \tag{6.8}$$

π can be identified as the z component of the vector potential (the vector potential has no other components).

We will assume π to be of the form

$$\pi = \hat{\pi}\,(x, y)e^{-\Gamma z} \tag{6.9}$$

Here $\hat{\pi}\,(x, y)$ is a function of x and y only, which specifies the field distribution in any x, y plane.

We can apply Maxwell's equations to obtain the electric fields

$$\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = j\omega\epsilon E_x$$

Using (6.7) and (6.8), and replacing differentiation with respect to z by multiplication by $-\Gamma$, we find

$$E_x = \frac{j\Gamma}{\omega\epsilon}\frac{\partial \pi}{\partial x} \tag{6.10}$$

Similarly

$$E_y = \frac{j\Gamma}{\omega\epsilon}\frac{\partial \pi}{\partial y} \tag{6.11}$$

We see that in an x, y plane, a plane perpendicular to the direction of propagation, the field is given as the gradient of a scalar potential V

$$V = (-j\Gamma/\omega\epsilon)\pi \tag{6.12}$$

This is because we deal with transverse magnetic waves, that is, with waves which have no longitudinal or z component of magnetic field. Thus, a closed path in an x, y plane, which is normal to the direction of propagation, will link no magnetic flux, and the integral of the electric field around such a path will be zero.

We can apply the curl relation and obtain E_z

$$\begin{aligned} \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= j\omega\epsilon E_z \\ E_z &= \frac{j}{\omega\epsilon}\left(\frac{\partial^2 \pi}{\partial x^2} + \frac{\partial^2 \pi}{\partial y^2}\right) \end{aligned} \tag{6.14}$$

Applying Maxwell's equations again, we have

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = j\omega\mu H_x$$
$$\frac{j}{\omega\epsilon}\frac{\partial}{\partial y}\left(\frac{\partial^2\hat{\pi}}{\partial x^2} + \frac{\partial^2\hat{\pi}}{\partial y^2}\right) + \frac{j\Gamma^2}{\omega\epsilon}\frac{\partial\hat{\pi}}{\partial y} = -j\omega\mu\frac{\partial\hat{\pi}}{\partial y} \tag{6.15}$$

This is certainly true if

$$\frac{\partial^2\hat{\pi}}{\partial x^2} + \frac{\partial^2\pi}{\partial y^2} = -(\Gamma^2 + \beta_0^2)\hat{\pi} \tag{6.16}$$

$$\beta_0 = \omega\sqrt{\mu\epsilon} = \omega/c \tag{6.17}$$

We find that this satisfies the other curl E relations as well.

From (6.16) and (6.14) we see that

$$E_z = (-j/\omega\epsilon)(\Gamma^2 + \beta_0^2)\hat{\pi}(x, y)e^{-\Gamma z} \tag{6.18}$$

For a given physical circuit, it will be found that there are certain real functions $\hat{\pi}_n(x, y)$ which are zero over the conducting boundaries of the circuit, assuring zero tangential field at the surface of the conductor, and which satisfy (6.16) with some particular value of Γ, which we will call Γ_n. Thus, as a particular example, for a square waveguide of width W some (but not all) of these functions are

$$\hat{\pi}_n(x, y) = \cos(n\pi y/W)\cos(n\pi x/W) \tag{6.19}$$

where n is an integer. We see from (6.10), (6.11) and (6.18) that this makes E_x, E_y and E_z zero at the conducting walls $x = \pm W/2$, $y = \pm W/2$.

Each possible real function $\hat{\pi}_n(x, y)$ is associated with two values of Γ_n, one the negative of the other. The Γ_n's are the natural propagation constants of the normal modes, and the $\hat{\pi}_n$'s are the functions giving their field distribution in the x, y plane. The $\hat{\pi}_n$'s can be shown to be orthogonal, at least in typical cases. That is, integrating over the region in the x, y plane in which there is field

$$\int\int \hat{\pi}_n(x, y)\,\hat{\pi}_m(x, y)\,dx\,dy = 0$$
$$n \neq m \tag{6.20}$$

For a lossless circuit the various field distributions fall into two classes: those for which Γ_n is imaginary, called active modes, which represent waves which propagate without attenuation; and those for which Γ_n is real, which change exponentially with amplitude in the z direction but do not change in phase. The latter can be used to represent the disturbance in a waveguide below cutoff frequency, for instance.

If Γ_n is imaginary (an active mode) the power flow is real, while if Γ_n is real (a passive mode) the power flow is imaginary (reactive or "wattless" power).

The spatial distribution functions $\hat{\pi}_n$ and the corresponding propagation constants Γ_n are a means for specifying the electrical properties of a physical structure, just as are the physical dimensions which describe the physical structure and determine the various $\hat{\pi}_n$'s and Γ_n's. In fact, if we know the various π_n's and Γ_n's, we can determine the response of the structure to an impressed current without direct reference to the physical dimensions.

In terms of the $\hat{\pi}_n$'s and Γ_n's, we can represent any unforced disturbance in the circuit in the form

$$\sum_n \hat{\pi}_n(x, y)[A_n e^{-\Gamma_n z} + B_n e^{\Gamma_n z}] \tag{6.21}$$

Here A_n is the complex amplitude of the wave of the nth spatial distribution traveling to the right, and B_n the complex amplitude of the wave of the same spatial distribution traveling to the left.

It is of interest to consider the power flow in terms of the amplitude, A_n or B_n. We can obtain the power flow P by integrating the Poynting vector over the part of the x, y plane within the conducting boundaries

$$\begin{aligned} P &= \frac{1}{2}\int\int EXH^* \, ds \\ P &= \frac{1}{2}\int\int (E_x H_y^* - E_y H_x^*) \, dx \, dy \end{aligned} \tag{6.22}$$

By expressing the fields in terms of the stream function, we obtain

$$P = A_n A_n^* \left(\frac{-j\Gamma_n}{2\omega\epsilon}\right) \iint \left[\left(\frac{\partial \hat{\pi}_n}{\partial x}\right)^2 + \left(\frac{\partial \hat{\pi}_n}{\partial y}\right)^2\right] dx \, dy \tag{6.23}$$

We can transform this by integrating by parts (essentially Green's theorem). Thus

$$\int_{x_1}^{x_2} \frac{\partial \hat{\pi}_n}{\partial x} \frac{\partial \hat{\pi}_n}{\partial x} \, dx = \hat{\pi}_n \frac{\partial \hat{\pi}_n}{\partial x}\Bigg|_{x_1}^{x_2} - \int_{x_1}^{x_2} \hat{\pi}_n \frac{\partial^2 \hat{\pi}_n}{\partial x^2} \, dx \tag{6.24}$$

Here x_1 and x_2, the limits of integration, lie on the conducting boundaries where $\hat{\pi}_n = 0$, and hence the first term on the right is zero. Doing the same for the second term in (6.23), we obtain

$$P_n = A_n A_n^* \left(\frac{-j\Gamma_n}{2\omega\epsilon}\right) \iint \hat{\pi}_n \left(\frac{\partial^2 \hat{\pi}_n}{\partial x^2} + \frac{\partial^2 \hat{\pi}_n}{\partial y^2}\right) dx \, dy \tag{6.25}$$

By using (6.16), we obtain

$$P_n = A_n A_n^* \left(\frac{j\Gamma_n}{2\omega\epsilon}\right)(\Gamma_n^2 + \beta_0^2) \iint (\hat{\pi}_n)^2 \, dx \, dy \tag{6.26}$$

It is also of interest to express the z component of the nth mode, E_{zn}, explicitly. For the wave traveling to the right we have, from (6.18),

$$E_{zn} = A_n \left(\frac{-j}{\omega\epsilon}\right)(\Gamma_n^2 + \beta_0^2)\hat{\pi}_n(x, y) \tag{6.27}$$

Let the field at some particular position, say, $x = y = 0$, be E_{zn0}. Then

$$A_n = \frac{j\omega\epsilon E_{zn0}}{(\Gamma_n^2 + \beta_0^2)\hat{\pi}_n(0, 0)} \tag{6.28}$$

and from (6.26)

$$P_n = (E_{zn0} E_{zn0}^*) \frac{-j\omega\epsilon\Gamma_n}{2\pi_n^2(0, 0)(\Gamma_n^2 + \beta_0^2)} \iint [\hat{\pi}_n(x, y)]^2 \, dx \, dy \tag{6.29}$$

We can rewrite this

$$\frac{E_{zn0} E_{zn0}^*}{(-\Gamma_n^2)P_n} = \frac{2\hat{\pi}_n^2(0, 0)(\Gamma_n^2 + \beta_0^2)}{-j\omega\epsilon\Gamma_n(-\Gamma_n^2) \iint [\hat{\pi}_n(x, y)]^2 \, dx \, dy} \tag{6.30}$$

For an active mode in a lossless circuit, Γ_n is a pure imaginary, and the negative of its square is the square of the phase constant. Thus, for a particular mode of propagation we can identify (6.30) with the circuit parameter $E^2/\beta^2 P$ which we used in Chapter II.

Let us now imagine that there is an impressed current J which flows in the z direction and has the form

$$J = \hat{J}(x, y)e^{-\Gamma z} \tag{6.31}$$

According to Maxwell's equations we must have

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = j\omega\epsilon E_z + J \tag{6.32}$$

Now, we will assume that the fields are given by some overall stream function π which varies with x and y and with z as $\exp(-\Gamma z)$.

In terms of this function π, H_x, H_y and E_x, E_y will be given by relations (6.7), (6.8), (6.10), (6.11). However, the relation used in obtaining E_z is not valid in the presence of the convection current. Instead of (6.16) we have

$$\begin{aligned} \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= j\omega\epsilon E_z + J \\ E_z &= \frac{j}{\omega\epsilon}\left(\frac{\partial^2\pi}{\partial x^2} + \frac{\partial^2\pi}{\partial y^2}\right) + \frac{j}{\omega\epsilon} J \end{aligned} \tag{6.33}$$

Again applying the relation

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = -j\omega\mu H_x$$

we obtain

$$\frac{\partial^2 \pi}{\partial x^2} + \frac{\partial^2 \pi}{\partial y^2} = -(\Gamma^2 + \beta_0^2)\,\pi - J \tag{6.34}$$

We will now divide both π and J into the spatial distributions characteristic of the normal unforced modes.
Let

$$\hat{J}(x, y) = \sum_n J_n \hat{\pi}_n(x, y) \tag{6.35}$$

$$J_n = \frac{\iint \hat{J}(x, y)\hat{\pi}_n(x, y)\,dx\,dy}{\iint [\hat{\pi}_n(x, y)]^2\,dx\,dy} \tag{6.36}$$

This expansion is possible because the π_n's are orthogonal. Let

$$\hat{\pi} = e^{-\Gamma z} \sum_n C_n \hat{\pi}_n(x, y) \tag{6.37}$$

Here there is no question of forward and backward waves; the forced excitation has the same z-distribution as the forcing current.

For the nth component, we have, from (6.16),

$$\frac{\partial^2 \hat{\pi}_n(x, y)}{\partial x^2} + \frac{\partial^2 \hat{\pi}_n(x, y)}{\partial y^2} = -(\Gamma_n^2 + \beta_0^2)\hat{\pi}_n(x, y) \tag{6.38}$$

From (6.34) we must also have

$$C_n\left(\frac{\partial^2 \hat{\pi}_n(x, y)}{\partial x^2} + \frac{\partial^2 \hat{\pi}_n(x, y)}{\partial y^2}\right) = -C_n(\Gamma^2 + \beta_0^2)\hat{\pi}_n(x, y) - J_n \hat{\pi}_n(x, y) \tag{6.39}$$

Accordingly, we must have

$$C_n = \frac{J_n}{\Gamma_n^2 - \Gamma^2} \tag{6.40}$$

The overall stream function is thus

$$\pi = e^{-\Gamma z} \sum_n \frac{\hat{\pi}_n(x, y) J_n}{\Gamma_n^2 - \Gamma^2} \tag{6.41}$$

From (6.33) and (6.34) we see that

$$E_z = \frac{-j}{\omega\epsilon}(\Gamma^2 + \beta_0^2)\pi \tag{6.42}$$

So

$$E_z = e^{-\Gamma z} \sum \frac{-j(\Gamma^2 + \beta_0^2)\hat{\pi}_n(x, y)J_n}{\omega\epsilon(\Gamma_n^2 - \Gamma^2)} \tag{6.43}$$

$$E_z = \frac{-j(\Gamma^2 + \beta_0^2)}{\omega\epsilon} e^{-\Gamma z} \sum \frac{\hat{\pi}_n(x, y)J_n}{\Gamma_n^2 - \Gamma^2} \tag{6.44}$$

6.2 Comparison with Results of Chapter II

Let us consider a case in which there is only one mode of propagation, characterized by $\hat{\pi}_1(x, y)$, Γ_1, and a case in which the current flows over a region in which $\hat{\pi}_1(x, y)$ has a constant value, say, $\hat{\pi}_1(0, 0)$. This corresponds to the case of the transmission line which was discussed in Chapter II.

We take only the term with the subscript 1 in (6.44) and (6.30). Combining these equations, we obtain for the field at 0, 0

$$E_z = \frac{(E^2/\beta^2 P)(\Gamma^2 + \beta_0^2)}{(\Gamma_1^2 + \beta_0^2)} \frac{\Gamma_1^3 J_1 \iint [\hat{\pi}_1(x, y)]^2 \, dx \, dy}{2\hat{\pi}_1(0, 0)} \tag{6.45}$$

We have from (6.36)

$$J_1 = \frac{\pi_1(0, 0)}{\iint [\hat{\pi}_1(x, y)]^2 \, dx \, dy} \tag{6.46}$$

From (6.45) and (6.46) we obtain

$$E_z = \frac{(\Gamma^2 + \beta_0^2)\Gamma_1^3(E^2/\beta^2 P)}{2(\Gamma_1^2 + \beta_0^2)(\Gamma_1^2 - \Gamma^2)} J e^{-\Gamma z} \tag{6.47}$$

Let us compare this with (6.4), which came from the transmission line analogy of Chapter II, identifying E_z and J with E and i. We see that, for slow waves for which

$$\beta_0^2 \ll |\Gamma_1^2| \tag{6.48}$$

$$\beta_0^2 \ll |\Gamma^2| \tag{6.49}$$

(6.47) becomes the same as (6.4). It was, of course, under the assumption that the waves are slow that we obtained (2.10), which led to (6.4).

6.3 Expansion Rewritten in Another Form

Expression (6.44) can be rewritten so as to appear quite different. We can write

$$\Gamma^2 + \beta_0^2 = \Gamma^2 - \Gamma_n^2 + \Gamma_n^2 + \beta_0^2$$

Thus, we can rewrite the expression for E_z as

$$E_z = e^{-\Gamma z}\left((-j/\omega\epsilon)\sum_n \frac{(\Gamma_n^2 + \beta_0^2)\hat{\pi}_n(x, y)J_n}{\Gamma_n^2 - \Gamma^2} + (j/\omega\epsilon)\sum_n \hat{\pi}_n(x, y)J_n\right) \tag{6.50}$$

The second term in the brackets is just $j/\omega\epsilon$ times the impressed current, as we can see from (6.35). The first term can be rearranged

$$(-j/\omega\epsilon)(\Gamma_n^2 + \beta_0^2)J_n = \frac{(-j/\omega\epsilon)(\Gamma_n^2 + \beta_0^2)\iint \hat{\pi}_n(x, y)J(x, y)\,dx\,dy}{\iint [\hat{\pi}_n(x, y)]^2\,dx\,dy} \tag{6.51}$$

Referring back to (6.29), let Ψ_n be twice the power P_n carried by the unforced mode when the field strength is

$$|E_{zn0}| = 1 \tag{6.52}$$

Further, let us choose the $\hat{\pi}_n$'s so that, at some specified position, $x = y = 0$,

$$_n(0, 0) = 1 \tag{6.53}$$

Then

$$\Psi_n = \frac{-j\omega\epsilon\Gamma_n}{\Gamma_n^2 + \beta_0^2}\iint [\hat{\pi}_n(x, y)]^2\,dx\,dy \tag{6.54}$$

Using this in connection with (6.51), we obtain

$$E_z = e^{-\Gamma z}\left(-\sum_n \frac{\Gamma_n\hat{\pi}_n(x, y)\iint \hat{\pi}_n(x, y)\hat{J}(x, y)\,dx\,dy}{\Psi_n(\Gamma_n^2 - \Gamma^2)} + (j/\omega\epsilon)\hat{J}(x, y)\right) \tag{6.55}$$

An expression for the forced field in terms of the parameters of the normal modes was given earlier[1,2]. In deriving this expression, the existence of a set of modes was assumed, and the field at a point was found as an integral over the disturbances induced in the circuit to the right and to the left and propagated to the point in question. Such a derivation applies for lossy and mixed waves, while that given here applies for lossless transverse-magnetic waves only.

The earlier derivation[1] leads to an expression identical with (6.55) except that Ψ_n^* appears in place of Ψ_n. In this earlier derivation a sign was implicitly assigned to the direction of flow of reactive power (which really doesn't flow at all!) by saying that the reactive power flows in the direction in which the amplitude decreases. If we had assumed the reactive power to flow in the direction in which the amplitude increases, then, with the same definition of Ψ_n, for a passive mode Ψ_n^* would have been replaced by $-\Psi_n^*$ which is equal to Ψ_n (for a passive mode, Ψ_n is imaginary).

In deriving (6.55), no such ambiguity arose, because the power flow was identified with the complex Poynting vector for the particular type of wave considered. In any practical sense, Ψ is merely a parameter of the circuit, and it does not matter whether we call Im Ψ reactive power flow to the right or to the left.

The existence of a derivation of (6.55) not limited in its application to lossless transverse magnetic waves is valuable in that practical circuits often have some loss and often (in the case of the helix, for instance) propagate mixed waves.

6.4 Iterated Structures

Many circuits, such as those discussed in Chapter IV, have structure in the z direction. Expansions such as (6.55) do not strictly apply to such structures. We can make a plausible argument that they will be at least useful if all field components except one differ markedly in propagation constant from the impressed current. In this case we save the one component which is nearly in synchronism with the impressed current and hope for the best.

APPENDIX III

STORED ENERGIES OF CIRCUIT STRUCTURES

A3.1 Forced Sinusoidal Field

If $v \ll c$, the field can be very nearly represented inside the cylinder of radius a by

$$V = V_0 \frac{I_0(\beta r)}{I_0(\beta a)} e^{-j\beta z} = \frac{E}{j\beta} \frac{I_0(\beta r)}{I_0(\beta a)} e^{-j\beta z} \tag{1}$$

and outside by

$$V = V_0 \frac{K_0(\gamma r)}{K(\gamma a)} e^{-j\beta z} \tag{2}$$

Inside

$$\frac{\partial V}{\partial r} = \beta \frac{I_1(\beta r)}{I_0(\beta a)} e^{-j\beta z} V_0 \tag{3}$$

$$\frac{\partial V}{\partial z} = -j\beta \frac{I_0(\beta r)}{I_0(\beta a)} e^{-j\beta z} V_0 \tag{4}$$

Outside

$$\frac{\partial V}{\partial r} = -\beta \frac{K_1(\beta r)}{K_0(\beta a)} e^{-j\beta z} V_0 \tag{5}$$

$$\frac{\partial V}{\partial z} = -j\beta \frac{K_0(\beta r)}{K_0(\beta a)} e^{-j\beta z} V_0 \tag{6}$$

Because there is a sinusoidal variation in the z direction, the average stored electric energy per unit length will be

$$W_E = \left(\frac{1}{2}\right)\left(\frac{\epsilon}{2}\right) \int_{r=0}^{\infty} [(E_{r\,\max})^2 + (E_{z\,\max})^2](2\pi r\, dr) \tag{7}$$

Here $E_{r\ max}$ and $E_{z\ max}$ are maximum values at $r = a$. The total electric plus magnetic stored energy will be twice this. This gives

$$W = \frac{\pi\epsilon(\gamma a)^2}{2\gamma^2}\left[\frac{I_0^2 - I_0 I_2}{I_0^2} + \frac{K_0 K_2 - K_0^2}{K_0^2}\right] E^2$$
$$W = \frac{\pi\epsilon\gamma a}{\gamma^2}\left[\frac{I_1}{I_0} + \frac{K_1}{K_0}\right] E^2 \tag{8}$$

$$(E^2/\beta^2 P)^{1/3} = (c/v)^{1/3}(v/v_g)^{1/3}\left[\frac{120}{\beta a\left(\frac{I_1}{I_0} + \frac{K_1}{K_0}\right)}\right]^{1/3} \tag{9}$$

A3.2 Pill-Box Resonators

Schelkunoff gives on page 268 of Electromagnetic Waves an expression for the peak electric energy stored in a pill-box resonator, which may be written as

$$.135\ \pi\ \epsilon\ a^2 h E^2$$

Here a is the radius of the resonator and h is the axial length. For a series of such resonators, the peak stored electric energy per unit length, which is also the average electric plus magnetic energy per unit length, is

$$W = .135\ \pi\ \epsilon\ a^2 E^2 \tag{10}$$

For resonance

$$a = 1.2\lambda_0/\pi \tag{11}$$

Whence

$$W = .0618\ \epsilon\lambda_0^2 E^2 \tag{12}$$

And

$$(E^2/\beta^2 P)^{1/3} = 5.36\ (v/v_g)^{1/3}\ (v/c)^{1/3} \tag{13}$$

The case of square resonators is easily worked out.

A3.3 Parallel Wires

Let us consider very fine very closely spaced half-wave parallel wires with perpendicular end plates.

If z is measured along the wires, and y perpendicular to z and to the direction of propagation, the field is assumed to be

$$E_x = E\cos\beta x e^{\pm\beta y}\cos\frac{2\pi}{\lambda_0} z$$
$$E_y = E\sin\beta x e^{\pm\beta y}\cos\frac{2\pi}{\lambda_0} z \tag{14}$$

Here the + sign applies for $y < 0$ and the − sign for $y > 0$. We will then find that

$$W = 2W_E = \frac{\epsilon E^2 \lambda_0}{2} \int_0^\infty e^{-2\beta y}\, dy$$
$$W = \frac{\epsilon \lambda_0}{4\beta} E^2 \tag{15}$$

and

$$(E^2/\beta^2 P)^{1/3} = 6.20\ (v/v_g)^{1/3} \tag{16}$$

The surface charge density σ on one side of the array of wires (say, $y > 0$) is given by the y component of field at $y = 0$.

$$\sigma = \epsilon E_y = \epsilon E \sin \beta x \cos \frac{2\pi}{\lambda_0} z \tag{17}$$

This is related to the current I (flowing in the z direction) per unit distance in the x direction by

$$\frac{\partial I}{\partial z} = -\frac{\partial \sigma}{\partial t} \tag{18}$$

From (18) and (17) we obtain for the current on one side of the array

$$I = -\frac{j\omega\lambda_0 \epsilon}{2\pi} E \sin \beta x \sin \frac{2\pi}{\lambda_0} z \tag{19}$$

If we use the fact that $\omega\lambda_0/2\pi = c$ and $c\,\epsilon = 1/\sqrt{\mu/\epsilon}$, we obtain

$$I = \frac{-jE}{\sqrt{\mu/\epsilon}} \sin \beta x \sin \frac{2\pi}{\lambda_0} z \tag{20}$$

If R is the surface resistivity of either side ($y > 0$, $y < 0$) of the wires, when the wires act as a resonator (a standing wave) the average power lost per unit length for both sides is

$$P = \tfrac{1}{8} R \lambda_0 E^2/(\mu/\epsilon) \tag{21}$$

In this case the stored electric energy is half the value given by (15), and we find

$$Q = (\sqrt{\mu/\epsilon}/R)\ (v/c) \tag{22}$$

Factors Affecting Magnetic Quality*

By R. M. BOZORTH

IN THE preparation of magnetic materials for practical use it is important to know how to obtain products of the best quality and uniformity. In the scientific study of magnetism the goal is to understand the relation between the structure and composition on the one hand and the magnetic properties on the other. From both standpoints it is necessary to know the principal factors which influence magnetic behavior. These are briefly reviewed here.

The properties depend on chemical composition, fabrication and heat-treatment. Some properties, such as saturation magnetization, change only slowly with chemical composition and are usually unaffected by fabrication or heat treatment. On the contrary, permeability, coercive force and hysteresis loss are highly sensitive and show changes which are extreme among all the physical properties. Properties may thus be divided into *structure-sensitive* and *structure-insensitive* groups. As an example, Fig. 1 shows magnetization curves of permalloy after it has been (a) cold rolled, (b) annealed and cooled slowly, and (c) annealed and cooled rapidly. The maximum permeability varies with the treatment over a range of about 20 fold, while the saturation induction is the same within a few per cent. Structure sensitive properties such as permeability depend on small irregularities in atomic spacings, which have little effect on properties such as saturation induction.

Some of the more common sensitive and insensitive properties are listed in Table I. The principal physical and chemical factors which affect these properties are listed in column 3. Their various effects will now be briefly discussed and illustrated.

Phase Diagram

Some of the most drastic changes in properties occur when the fabrication or heat treatment has brought about a change in structure of the material. For this reason the phase diagram or constitutional diagram is of the utmost importance in relation to the preparation and properties of magnetic materials. As an example consider the phase diagram of the binary iron-cobalt alloys of Fig. 2. Here the various areas show the phases, of different

*This article is the substance of Chapter II of a book entitled "Ferromagnetism" to be published early in 1951 by D. Van Nostrand Company, Inc.

composition or structure, which are stable at the temperatures and compositions indicated. The α phase has the body-centered-cubic crystal structure characteristic of iron. At 910°C it transforms into the face-centered phase γ, and at 1400° into the δ phase, which has the same structure as the α phase. At about 400°C cobalt transforms, on heating, from the ϵ phase (hexagonal structure) into the γ phase.

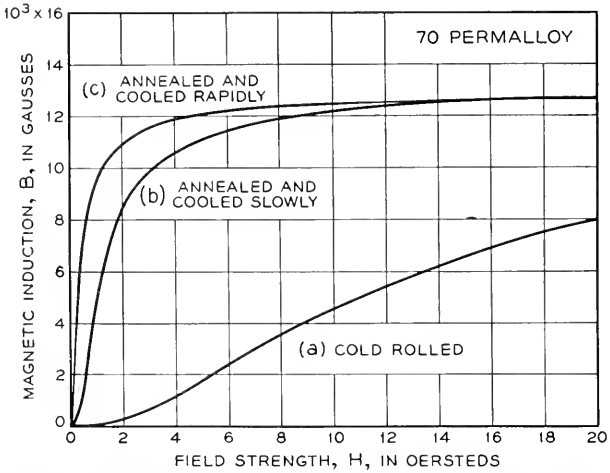

Fig. 1—Effect of mechanical and heat treatment on the magnetization curve of 70 permalloy (70% Ni, 30% Fe).

TABLE I

Properties Commonly Sensitive or Insensitive to Small Changes in Structure, and Some of the Factors which Effect Such Changes

Structure-Insensitive Properties	Structure-Sensitive Properties	Factors Affecting the Properties
I_s, Saturation Magnetization	μ, Permeability	Composition (gross)
θ, Curie Point	H_c Coercive Force	Impurities
λ_s, Magnetostriction at Saturation	W_h Hysteresis Loss	Strain
K, Crystal Anisotropy Constant		Temperature
		Crystal Structure
		Crystal Orientation

The dotted lines indicate the Curie point, at which the material becomes non-magnetic.

In between the areas corresponding to the single phases α, γ, δ and ϵ there are two-phase regions in which two crystal structures co-exist, some of the crystal grains having one structure and others the other. Such a two-phase structure is usually evident upon microscopic or X-ray examina-

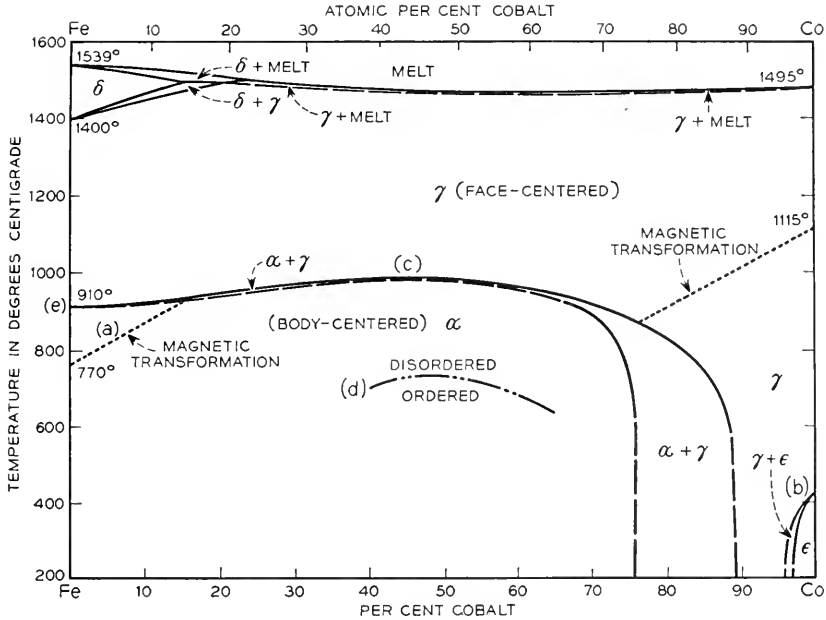

Fig. 2—Phase diagram of iron-cobalt alloys.

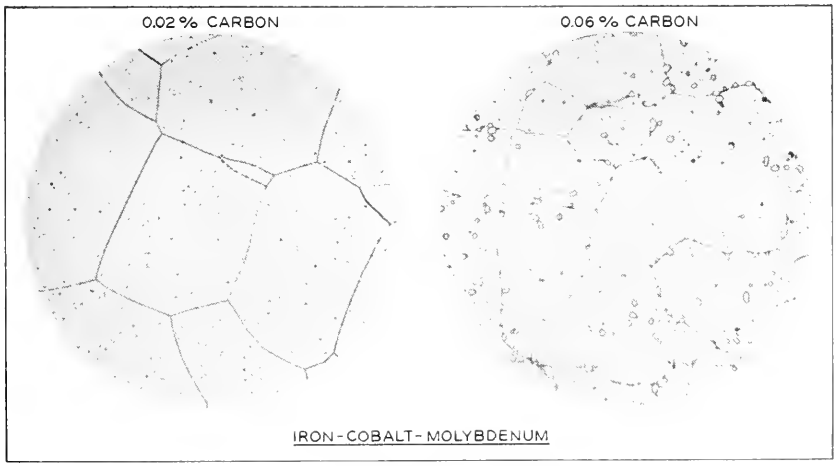

Fig. 3—Photomicrographs of remalloy (12% Co, 17% Mo, 71% Fe) showing the precipitation of a second phase in the specimen containing an excess of carbon (0.06% C) Courtesy of E. E. Thomas. Magnification: (a) 50 times, (b) 200 times.

tion. Microphotographs of a single-phase alloy and a two-phase alloy of iron-cobalt-molybdenum are reproduced in Fig. 3 (a) and (b).

The diagram of Fig. 2 shows several kinds of changes that affect the magnetic properties. At (a) the material becomes non-magnetic on heating, without change in phase. At (b) there is a change of phase, both phases

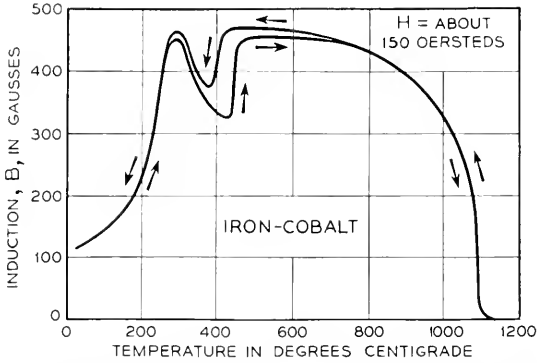

Fig. 4—Effect of phase transformation of cobalt on magnetization with a constant field of 150 oersteds. Both phases magnetic. *Masumoto.*

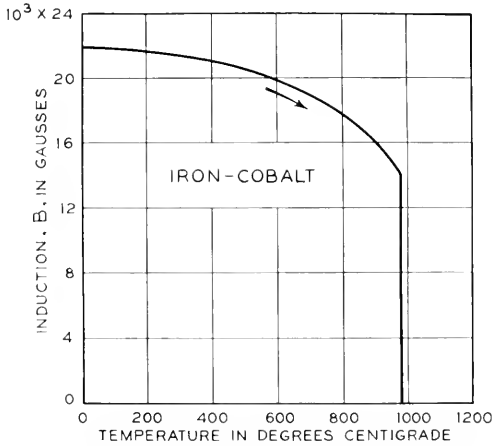

Fig. 5 - Phase transformation in iron-cobalt alloy (50% Co). High-temperature phase is non-magnetic.

being magnetic. Figure 4 shows the changes in magnetic properties that occur during this latter transition; they are due partly to the high local strains that result from the change in structure, and partly to the difference in the crystal structures of the two phases. At (c) there is a change from a ferromagnetic to a non-magnetic phase, and Fig. 5 shows the rapid change in magnetization that occurs when the temperature rises in this area. At

(d) the α phase becomes ordered on cooling, i.e., the iron and cobalt atoms tend to distribute themselves regularly among the various atom positions so that each atom is surrounded by atoms of the other kind. This phenomenon is especially important in connection with the properties of iron-aluminum and manganese-nickel alloys.

The transition at (e) is entirely in the non-magnetic region but it has its influence on the properties of iron at room temperature. If iron is cooled very slowly through (e), the internal strains caused by the change in structure will be relieved by diffusion of the metal atoms, but if the cooling is too rapid there will not be sufficient time for strain relief. Practically this means that to obtain high permeability in iron it must be annealed for some time below 900°C, or cooled slowly through this temperature so that diffusion will have time to occur. In most ferromagnetic materials diffusion occurs at a reasonably rapid rate only at temperatures above about 500 to 600°C.

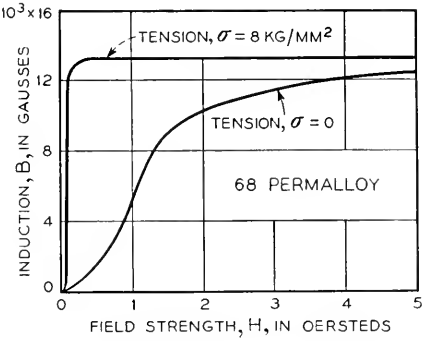

Fig. 6—Effect of tension on the magnetization curve of 68 permalloy.

The effect of a homogeneous *strain* on the magnetization curve can be observed in a simple way, as by applying tension to an annealed wire and then measuring B and H. The effect of tension on some materials is to increase the permeability and on other materials to decrease it, as shown in Fig. 6. Compression usually causes a change in the opposite sense.

The internal strains resulting from *plastic deformation* of the material, brought about by stressing beyond the elastic limit, as by pulling, rolling or drawing, almost always reduce the permeability. The material is then under rather severe local strains similar to those present after phase change, and these strains are different in magnitude and direction in different places in the material and have quite different values at points close together. Strains of this kind can usually be relieved by annealing; therefore, metal that has been fabricated by plastic deformation is customarily annealed to raise its permeability. Figure 1 shows the effect of annealing a permalloy strip that has been cold-rolled to 15 per cent of its original thickness.

The *temperature* also is effective in changing permeability and other properties, even when no change in phase occurs. Figure 7 shows the rapidity with which the initial permeability decreases as the Curie point is approached. For this material, Ferroxcube III, a zinc manganese ferrite ($ZnMnFe_4O_8$), the Curie point is not far above room temperature.

The effect of *impurities* may be illustrated by the B vs H curves for iron containing various amounts of carbon. Curve (a) of Fig. 8 is for a mild

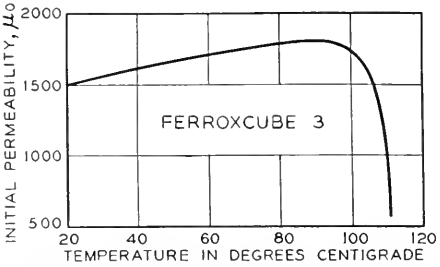

Fig. 7—Variation of initial permeability of Ferroxcube 3, showing maximum at temperature just below the Curie temperature.

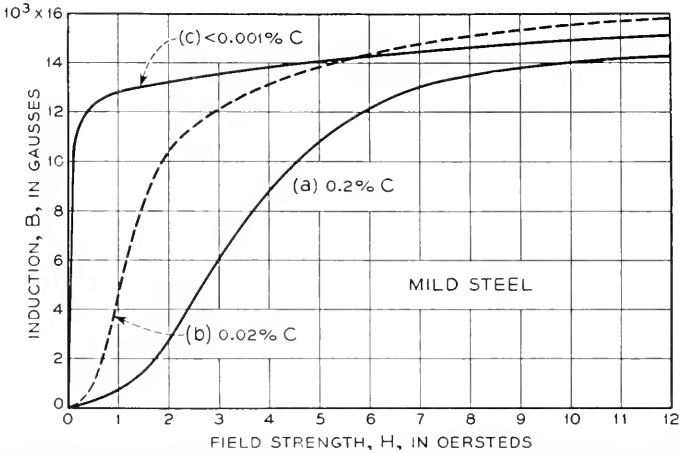

Fig. 8—Effect of impurities on magnetic properties of iron. Annealing at 1400°C in hydrogen reduces the carbon content from about 0.02 per cent to less than 0.001 per cent.

steel having 0.2 per cent carbon, (b) is for the iron commonly used in electromagnetic apparatus—it contains about 0.02 per cent carbon and is annealed at about 900°C. When this same iron is purified by heating for several hours at 1400°C in hydrogen, the carbon is reduced to less than 0.001 per cent and other impurities are removed, and curve (c) is obtained.

Finally, Fig. 9 shows that large differences in permeability may be found by simply varying the *direction of measurement* of the magnetic properties in a single specimen. The material is a single crystal of iron containing about 4 per cent silicon, and the directions in which the properties are measured

are [100] (parallel to one of the crystal axes), and [111] (as far removed as possible from an axis). The magnetic properties in the two directions are different because different "views" of the atomic arrangement are obtained in the two directions.

Production of Magnetic Materials

In the preparation of magnetic materials for either laboratory or commercial use there are many processes which influence the chemical and physical

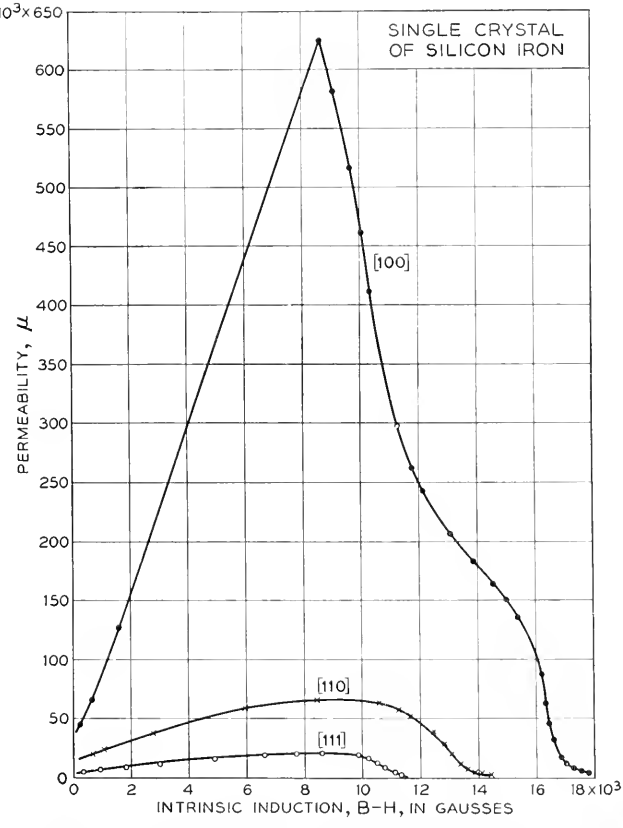

Fig. 9—Dependence of permeability on crystallographic direction. *Williams.*

structure of the product. The selection of raw materials, the melting and casting, the fabrication and the heat treatment, are all important and must be carried out with a proper knowledge of the metallurgy of the material. A brief description of the common practices is now given. For further discussion the reader is referred to more detailed metallurgical books and articles.

Melting and Casting

For experimental investigation of magnetic materials in the laboratory, the raw materials easily obtainable on the market are generally satisfactory. When high purity is desirable specially prepared materials and crucibles must be used and the atmosphere in contact with the melt must be controlled. The impurities that have the greatest influence on the magnetic properties of high permeability materials are the non-metallic elements,

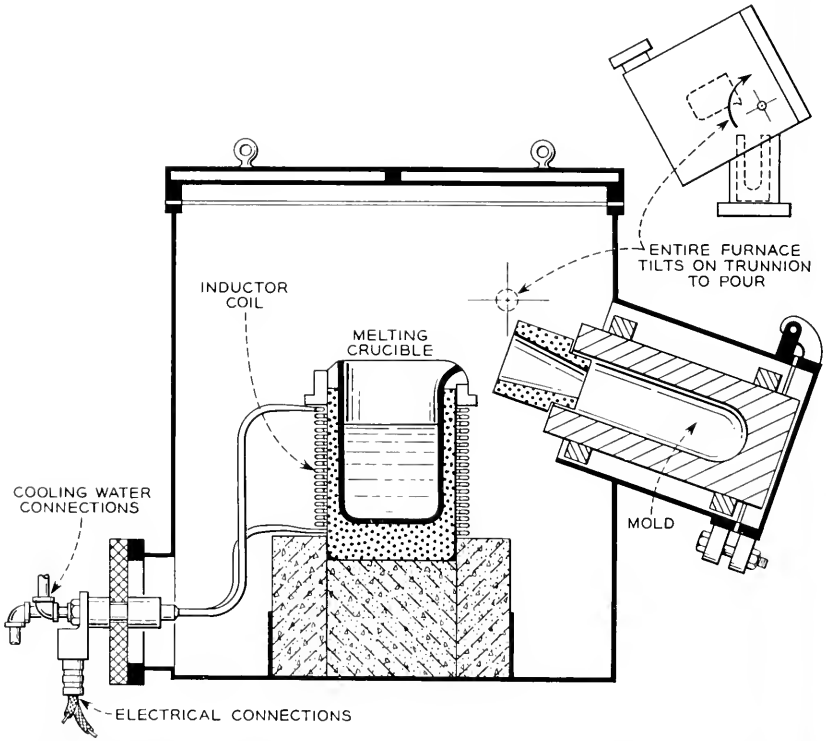

Fig. 10—Induction furnace designed for small melts in controlled atmosphere, as designed by J. H. Scaff and constructed by the Ajax Northrup Company.

particularly oxygen, carbon and sulfur, and the presence of these impurities is therefore watched carefully and their analyses are carried out with special accuracy. Impurities are likely to change in important respects during the melting and pouring on account of reactions of the melt with the atmosphere, the slag or the crucible lining, or because of reactions taking place among the constituents of the metal.

Melting of small lots (10 pounds) is best carried out in a high-frequency induction furnace. Figure 10 shows such a furnace designed for melting ten to fifty pounds, and casting by tilting the furnace, the whole operation being

carried out in a controlled atmosphere. High-frequency currents (usually 1,000 to 2,000 cycles/sec but sometimes much higher) are passed through the water-cooled copper coils, and the alternating magnetic field so produced

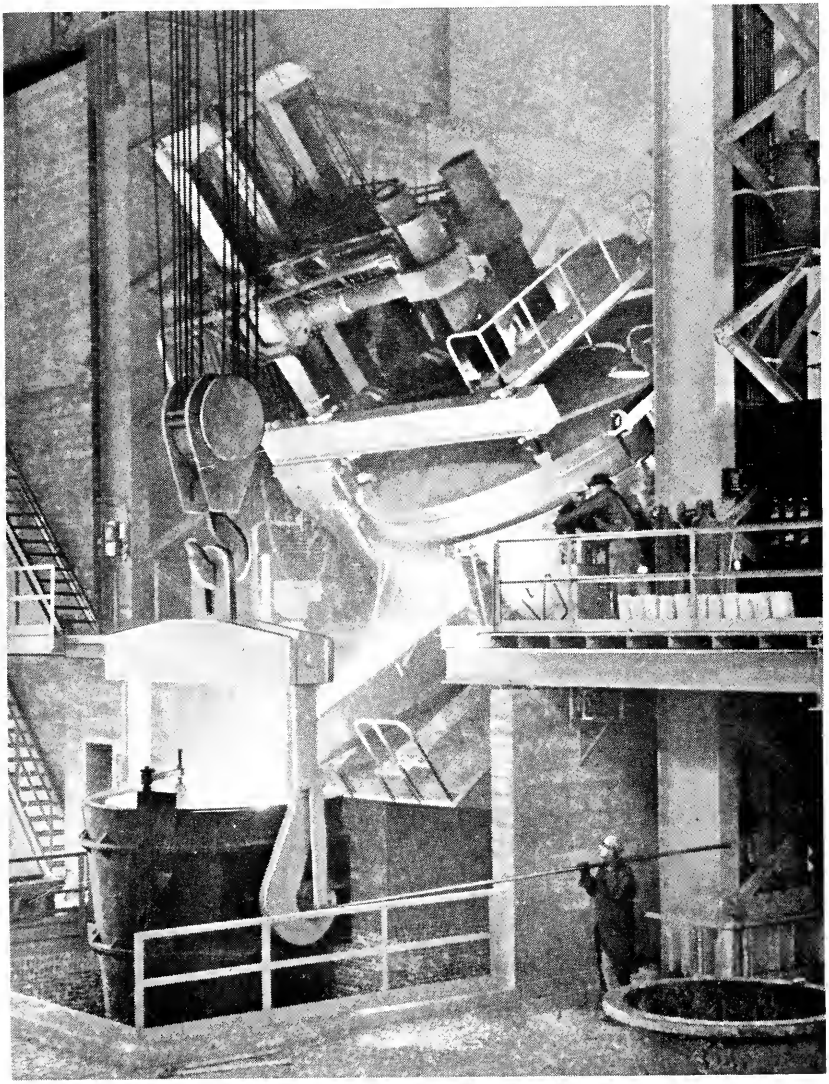

Fig. 11—Arc furnace for large commercial melts. Courtesy of J. S. Marsh of the Bethlehem Steel Company.

heats the charge by inducing eddy currents in it. Crucibles are usually composed of alumina or magnesia.

On a commercial scale melts of silicon-iron are usually made in the open

hearth furnace, in which pig-iron and scrap are refined and ferro-silicon added. The furnace capacity may be as large as 100 tons. Sometimes silicon-iron, and usually iron-nickel alloys, are melted in the arc furnace, in amounts varying from a few tons to 50 tons. A photograph of such a furnace, in the position of pouring, is shown in Fig. 11. The heat is produced in the arc drawn between large carbon electrodes immersed in the metal, the current sometimes rising to over 10,000 amperes. By tipping the furnace the melt is poured into a ladle, and from this it is poured into cast-iron molds through a valve-controlled hole in the ladle bottom. Special-purpose alloys, including permanent magnets, are prepared commercially in high-

TABLE II

Heats of Formation and Other Properties of Some Oxides (Sachs and Van Horn[4])

Oxide	Heat of formation (Kilo-cal per gram atom of metal)	Melting Point (°C)	Density (g/cm³)
CaO	152	>2500	3.4
BeO	144	>2500	3.0
MgO	144	2800	3.65
Li_2O	141	>1700	2.0
Al_2O_3	127	2050	3.5
V_2O_2	116	1970	4.9
TiO_2	109	1640	4.3
Na_2O	101	*	2.3
SiO_2	95	1670	2.3
B_2O_3	94	580	1.8
MnO	91	1650	5.5
ZrO_2	89	2700	5.5
ZnO	85	*	5.5
P_2O_5	73	*	2.4
SnO_2	68	1130	6.95
FeO	66	1420	5.7
NiO	58	**	7.45

* Sublimes.
** Decomposes before melting.

frequency induction furnaces or in arc furnaces in quantities ranging from a fraction of a ton to several tons.

Slags are commonly used when melting in air, both to protect from oxidation and to reduce the amounts of undesirable impurities. Common protective coverings are mixtures of lime, magnesia, silica, fluorite, alumina, and borax in varying proportions. In commercial production different slags are used at different stages, to refine the melt; e.g., iron oxide may be used to decarburize and basic oxides to desulfurize.

Melting in vacuum requires special technique that has been described in some detail by Yensen.[1] Commercial use has been described by Rohn[2] and others.[3] Melting in hydrogen has been used on an experimental scale in both

[1] T. D. Yensen, *Trans. A.I.E.E.* 34, 2601–41 (1915).
[2] W. Rohn, Heraeus Vacuumschmelze, Albertis, Hanau, 356–80 (1933).
[3] W. Hessenbruch and K. Schichtel, *Zeits. f. Metallkunde* 36, 127–30 (1944).

high-frequency and resistance-wound furnaces. In commercial furnaces Rohn has used hydrogen and vacuum alternately before pouring, for purification in the melt, in low-frequency induction furnaces having capacities of several tons.

Just before casting a melt of a high-permeability alloy such as iron nickel, a deoxidizer may be added, e.g. aluminum, magnesium, calcium or silicon, in an amount averaging around 0.1 per cent. The efficacy of a deoxidizer is measured by its heat of formation, and this is given for the common ele-

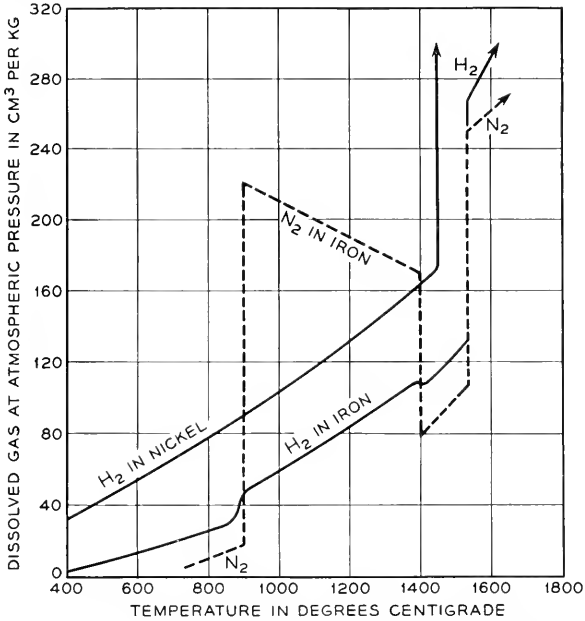

Fig. 12—Solubility of some gases in iron and nickel at various temperatures. *Sieverts.*

ments in Table II, taken from Sachs and Van Horn.[4] Also several tenths of a per cent of manganese may be put in to counteract the sulfur so that the material may be more readily worked; the manganese sulfide so formed collects into small globular masses which do not interfere seriously with the magnetic or mechanical properties of most materials.

Ordinarily a quantity of gas is dissolved in molten metal, and this is likely to separate during solidification and cause unsound ingots. The solubilities of some gases in iron and nickel have been determined by Sieverts[5] and others and are given in Fig. 12, adapted from the compilation by Dushman.[6] The characteristic decrease of solubility during freezing is apparent. Most

[4] G. Sachs and K. R. Van Horn, Practical Metallurgy, Am. Soc. Metals, Cleveland (1940).
[5] A. Sieverts, *Zeits. f. Metallkunde* 21, 37–46 (1929).
[6] S. Dushman, Vacuum Technique, Wiley, New York (1949).

of the gases given off by magnetic metals during heating are formed from the impurities carbon, oxygen, nitrogen and sulfur; CO is usually given off in greatest amount from cast metal, and some N_2 and H_2 are also found. Refining of the melt is therefore of obvious advantage, and the furnace of Fig. 10 is especially useful for this purpose.

Small ingots are sometimes made by cooling in the crucible. Usually, however, ingots are poured into cast iron molds for subsequent reduction by rolling, etc.; permanent magnet or other materials are often cast in sand

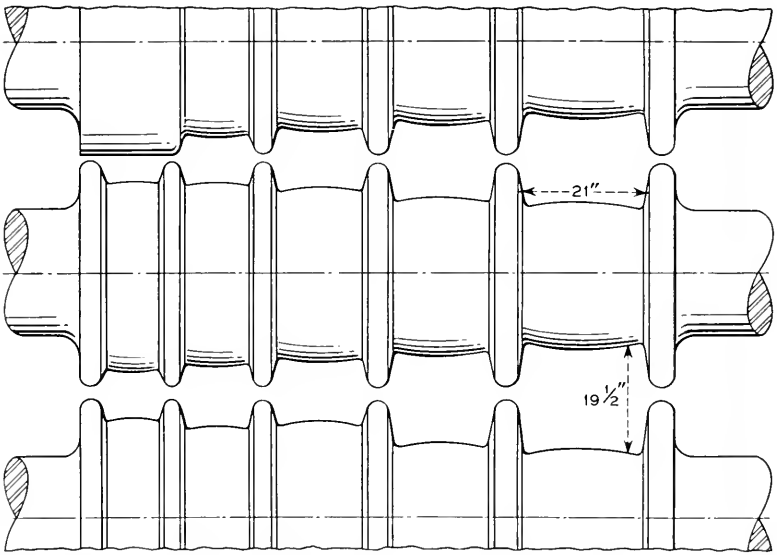

Fig. 13—Design of rolls in a blooming mill for hot reduction of ingots to rod. *Carnegie Illinois Steel Corp.*

in shapes which require only nominal amounts of machining or grinding for use in apparatus or in testing. Special techniques are used for specific materials.

Other considerations important in the melting and pouring of ingots are proper mixing in the melt, the temperature of pouring, mold construction, inclusions of slag, segregation, shrinkage, cracks, blow holes, etc.

Fabrication

Magnetic materials require a wide variety of modes of fabrication, which can best be discussed in connection with the specific materials. The methods include hot and cold rolling, forging, swaging, drawing, pulverization, elec-

trodeposition, and numerous operations such as punching, pressing and spinning. In the commercial fabrication of ductile material it is common practice to start the reduction in a breakdown or blooming mill (Fig. 13) after heating the ingot to a high temperature (1200° to 1400°C). Large ingots, of several tons weight, are often led to the mill before they have cooled below the proper temperature. The reduction is continued as the metal cools, in a rod or flat rolling mill, depending on the desired form of the final product. When the thickness is decreased to 0.2 to 0.5 inch the material has usually cooled below the recrystallization temperature. Because of the difficulty in handling hot sheets or rod of small thickness, they are rolled at or near room temperature, with intermediate annealings if necessary to soften or to develop the proper structure. In experimental work, rod is often swaged instead of rolled.

In recent years the outstanding trends in methods of fabricating materials have been toward the construction of the multiple-roll rolling mill for rolling thin strip, and the continuous strip mill for high-speed production on a large scale. Figure 14 shows the principle of construction of a typical 4-high mill ((a) and (b)), and of two special mills ((c) and (d)). In the 20-high Rohn[7] mill and 12-high Sendzimir[8] mill the two working rolls are quite small (0.2 to one inch in diameter). These are each backed by two larger rolls and these in turn by others as indicated. In the Rohn mill (c), power is supplied to the two smallest rolls and the final bearing surfaces are at the ends of the largest rolls. In the Sendzimir mill (d) the power is supplied to the rolls of intermediate size and the bearing surfaces are distributed along the whole length of the largest rolls so that no appreciable bending of the rolls occurs. The small rolls reduce the thickness of thin stock with great efficiency, and the idling rolls permit the application of high pressure. In the Steckel mill power is used to pull the sheet through the rolls, which are usually 4-high with small working rolls.

The continuous strip mill is an arrangement of individual mills such that the strip is fed continuously from one to another and may be undergoing reduction in thickness in several mills simultaneously. Figure 15 shows a mill of this kind, used for cold reduction, with 6 individual mills in tandem.

For magnetic testing numerous forms of specimens are required for various kinds of tests; these include strips for standard tests for transformer sheet, rings or parallelograms for conventional ballistic tests, "pancakes" of thin tape spirally wound for measurement by alternating current, ellipsoids for high field measurements, and many others. The various forms are

[7] W. Rohn, Heraeus Vacuumschmelze, Albertis, Hanau, 381–7 (1933).
[8] T. Sendzimir, *Iron and Steel Engr.* 23, 53–9 (1946).

required to study or eliminate the effects of eddy-currents, demagnetizing fields and directional effects and to simulate the use of material in apparatus. Most of the needs arizing in commerce and in experimental investigation are filled by strips or sheets of thicknesses from 0.002 inch to 0.1 inch from

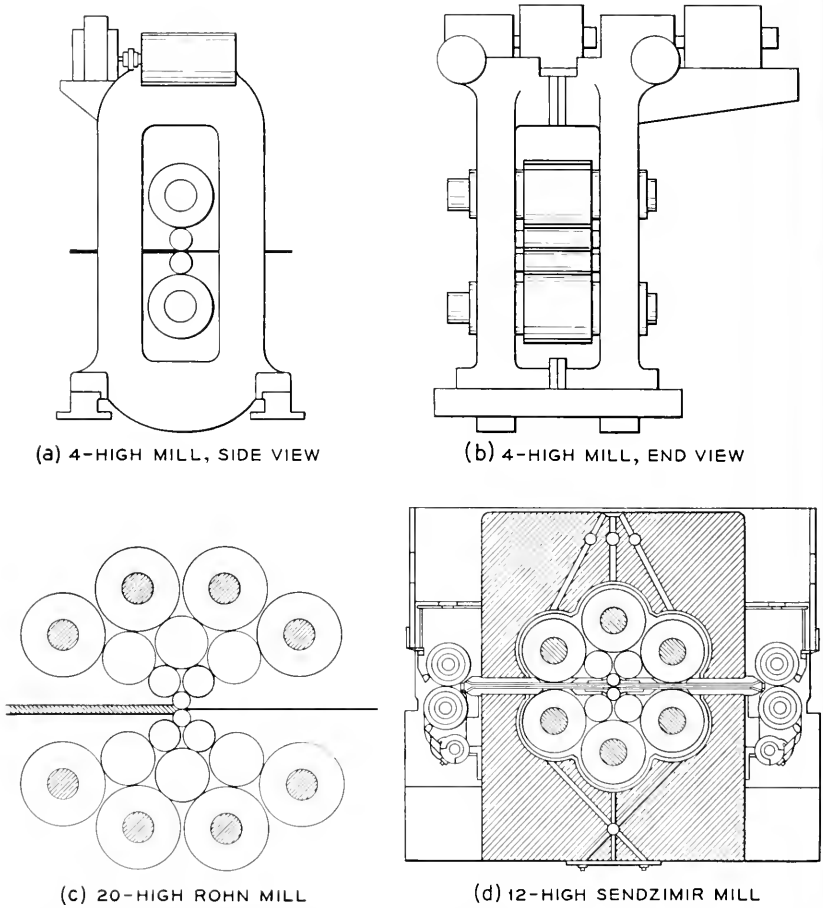

Fig. 14 —Arrangement of rolls in mills used for reduction of thin sheet: (a) and (b) conventional 4 high mill; (c) Rohn 20-high; (d) Sendzimir 12-high.

which coils can be wound or parts cut, by rods from which relay cores or other forms can be made, by powdered material used for pressing into cores for coils for inductive loading, and by castings for permanent magnets or other objects which may be machined or ground to final shape.

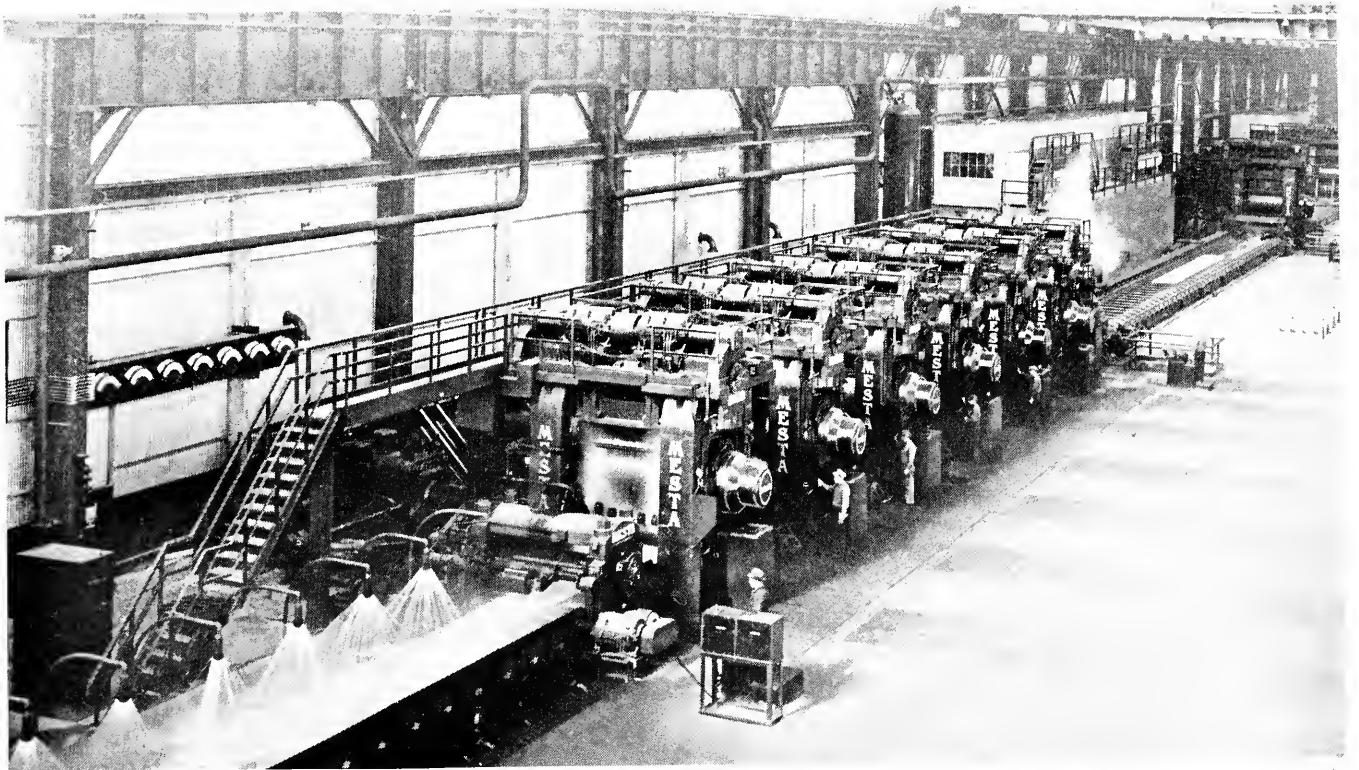

Fig. 15—Continuous strip mill designed for large output, having 6 individual mills in tandem. Courtesy of C. W. Stoker of Carnegie Illinois Steel Corp.

Heat-Treatment

High permeability materials are annealed primarily to relieve the internal strains introduced during fabrication. On the contrary permanent magnet materials are heat-treated to *introduce* strains by precipitating a second phase. Heat-treatments are decidedly characteristic of the materials and their intended uses and are best discussed in detail in connection with them.

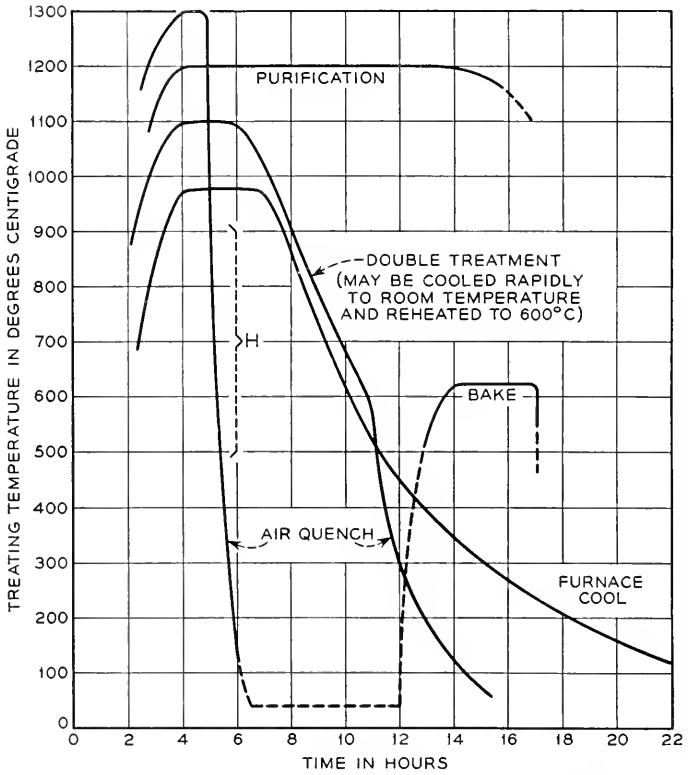

Fig. 16—Some common heat treatments for magnetic materials.

Figure 16 shows some of the commonest treatments in the form of temperature-time curves. The purpose of these various heating and cooling cycles, and typical materials subjected to them, may be listed as follows:

(1) Relief of internal strains due to fabrication or phase-changes (furnace cool). Magnetic iron.

(2) Increase of internal strains by precipitation hardening (air quench and bake). Alnico type of permanent magnets.

(3) Purification by contact with hydrogen or other gases. Silicon-iron (cold rolled), hydrogen-treated iron, Supermalloy.

There are also special treatments, such as those used for "double-treated" permalloy, "magnetically annealed" permalloy, and perminvar.

Occasionally it is necessary to homogenize a material by maintaining the

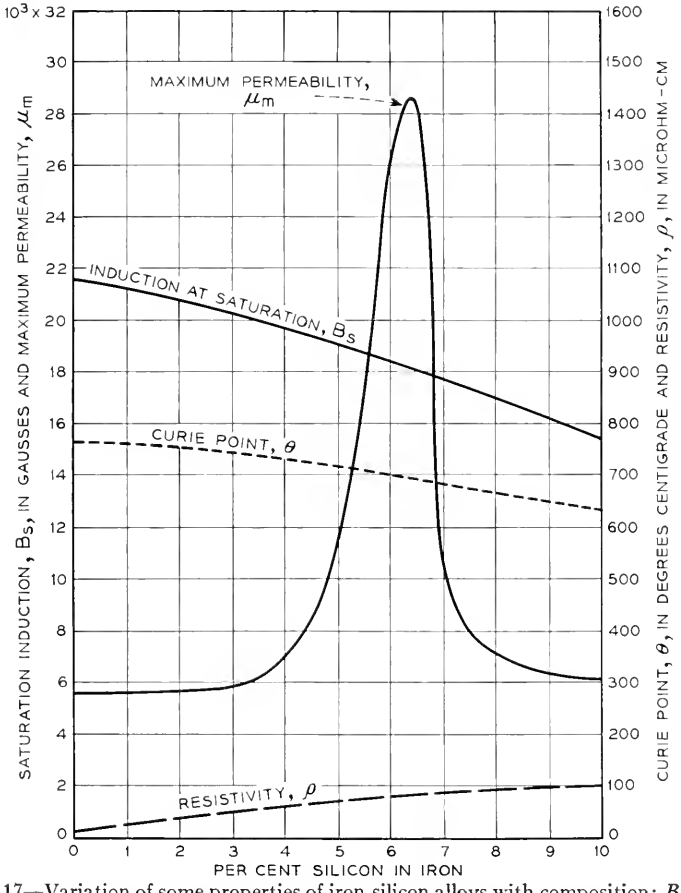

Fig. 17—Variation of some properties of iron-silicon alloys with composition: B_s, saturation intrinsic induction; θ, magnetic transformation point; ρ, electrical resistivity; μ_m, maximum permeability as determined by Miss M. Goertz.

temperature just below the freezing point for many hours. Heat-treatments also may affect grain size and crystal orientation.

Furnaces for heat-treating have various designs that will not be considered here. A modern improvement has been the use of globar (silicon carbide) heating elements that permit treatment at 1300 to 1350°C in an atmosphere of hydrogen or air.

Further discussion of "Metallurgy and Magnetism" is given in an excellent small book of this title by Stanley.[9]

Effect of Composition

Gross Chemical Composition

The effect of composition on magnetic properties will now be considered, using as examples the more important binary alloys of iron with silicon,

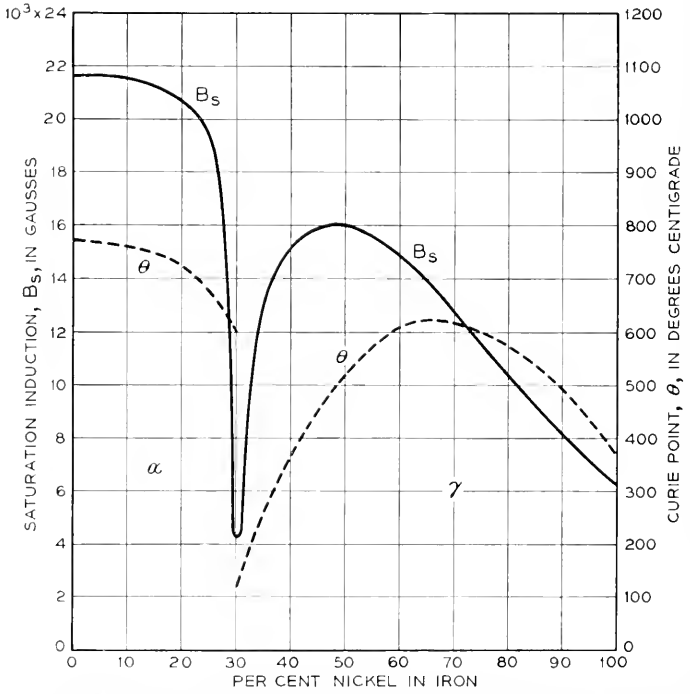

Fig. 18—Variation of B_s and θ with the composition of iron-nickel alloys.

nickel or cobalt, on which are based the most useful and interesting materials. The iron-silicon alloys are used commercially without additions, the iron-nickel and iron-cobalt alloys are most useful in the ternary form; and many special alloys, for example material for permanent magnets, contain four or five components.

Figure 17 shows four important properties of the iron-silicon alloys of low silicon content, after they have been hot rolled and annealed. The commercial alloys (3 to 5% silicon) are the most useful because they have the best

[9] J. K. Stanley, Metallurgy and Magnetism, Am. Soc. Metals, Cleveland (1949).

combination of properties of various kinds. The properties shown in the figure are important in determining the best balance: the maximum permeability, μ_m, only indirectly (it is a good measure of hysteresis loss and maximum field necessary in use), and the Curie point, θ, only in a minor role. The saturation B_s, permeability, and resistivity ρ, should all be as high as possible. B_s, θ and ρ are structure insensitive, and vary with composition in a characteristically smooth way, practically independent of heat treatment; μ_m depends on heat treatment (strain), impurities and crystal orientation. There are no phase changes to give sudden changes with composition of properties measured at room temperatures.

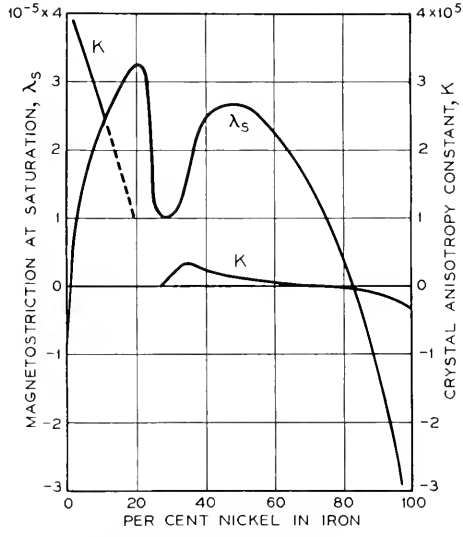

Fig. 19—Variation of saturation magnetostriction, λ_s, and crystal anisotropy, K, with the composition of iron-nickel alloys.

Some of the properties of the iron-nickel alloys are given in Figs. 18 and 19. The change in phase from α to γ at about 30 per cent nickel is responsible for the breaks at this composition. The permeabilities, μ_0 and μ_m, (Fig. 20) show characteristically the effect of heat treatment. The maxima are closely related to the points at which the saturation magnetostriction, λ_s, and crystal anisotropy, K, pass through zero (Fig. 19).

Additions of molybdenum, chromium, copper and other elements are made to enhance the desirable properties of the iron-nickel alloys.

The iron-cobalt alloys, some properties of which are shown in Fig. 21, are usually used when high inductions are advantageous. The unusual course of the saturation induction curve, with a maximum greater than that for any other material, is of obvious theoretical and practical importance. The sud-

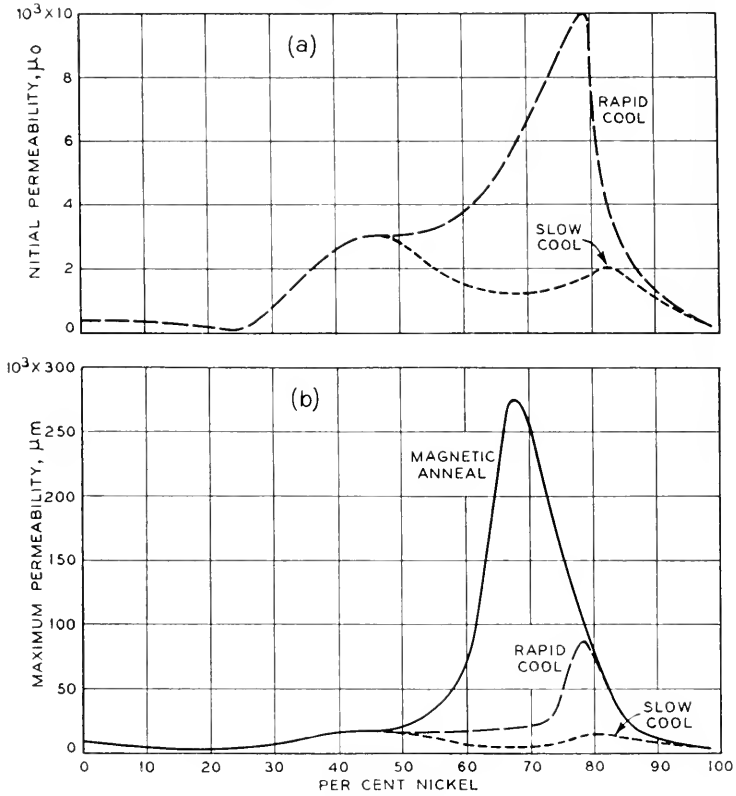

Fig. 20—Dependence of the initial and maximum permeabilities (μ_0, μ_m) of iron-nickel alloys on the heat treatment.

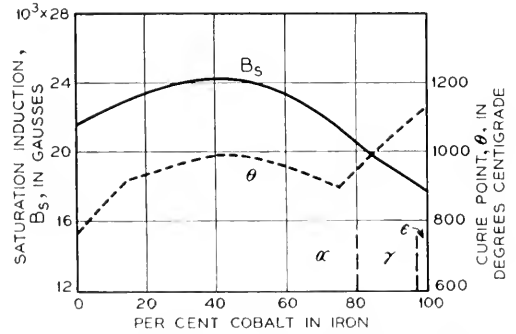

Fig. 21—Variation of B_s and θ of iron-cobalt alloys with composition.

den changes in the Curie point curve are associated with α, γ phase boundaries, as mentioned earlier in this chapter. The peak of the permeability curve (Fig. 22) occurs at the composition for which atomic ordering is stable at the highest temperature (see also Fig. 2). The sharp decline near 95 per cent cobalt coincides with the phase change γ,ϵ at this composition. Additions of vanadium, chromium and other elements are used in making commercial ternary alloys.

Some useful alloys based on the binary iron-silicon, iron-nickel and iron-cobalt alloys are described in Table III.

The *hardening* of material resulting from the precipitation of one phase in another is often used to advantage when magnetic hardness (as in permanent magnets), or mechanical hardness, is desired. To illustrate this

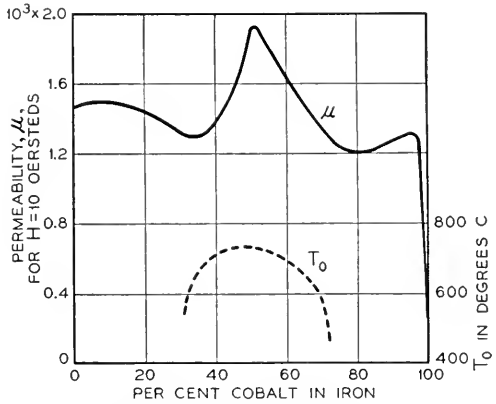

Fig. 22—Variation of permeability at $H = 10$ oersteds, and of the critical temperature of ordering, with the composition of iron-cobalt alloys.

process consider the binary iron molybdenum alloys, a partial phase diagram of which is given in Fig. 23. The effect of the boundary between the α and $\alpha + \epsilon$ fields is shown by the variation of the properties with composition (Fig. 24a). Saturation magnetization and Curie point are affected but little, the principle change in the former being a slight change in the slope of the curve at the composition at which the phase boundary crosses 500°C, the temperature below which diffusion is very slow. The Curie point curve has an almost imperceptible break at the composition at which the phase boundary lies at the Curie temperature. The changes of maximum permeability and coercive force are more drastic; μ_m drops rapidly as the amount of the second phase, ϵ, increases and produces more and more internal strain (Fig. 24b), and H_c increases at the same time. The experimental points correspond to a moderate rate of cooling of the alloy after annealing.

TABLE III

Some Properties of Some Useful Alloys Based on the Fe-Si, Fe-Co and Fe-Ni Binary Systems

Name	Composition (Per cent)	Heat Treatment	Initial Permeability, μ_0, μ	Maximum Permeability, μ_m	Coercive Force, H_c (oersteds)	Saturation Induction, B_s (gausses)	Curie Point, θ (°C)
Hot rolled Silicon Iron	4Si, 96Fe	800°C	500*	7000	0.5	19700	690
Grain Oriented Silicon Iron	3Si, 97Fe	1200°C	1500*	40000	0.15	20000	700
Sendust	9Si, 85Fe, 5Al	Cast	30000	120000	0.05	10000	500
45 Permalloy**	45Ni, 55Fe	1200°C, H_2	3500	50000	0.07	16000	440
4-79 Permalloy	79Ni, 17Fe, 4Mo	1100°C	20000	100000	0.05	8700	420
Mumetal	75Ni, 18Fe, 2Cr, 5Cu	1175°C, H_2	20000	100000	0.05	6500	430
Supermalloy	79Ni, 16Fe, 5Mo	1300°C, H_2	100000	1000000	0.002	8000	400
Permendur	50Co, 50Fe	800°C	800	5000	2.0	24500	980
2V-Permendur	49Co, 49Fe, 2V	800°C	800	4500	2.0	24000	980
Hiperco	34Co, 64Fe, 1Cr	850°C	650	10000	1.0	24200	—

* Measured at B = 20 instead of B = 0.

** Similar alloys: Hipenik, Nicaloi, 4750 alloy, and others.

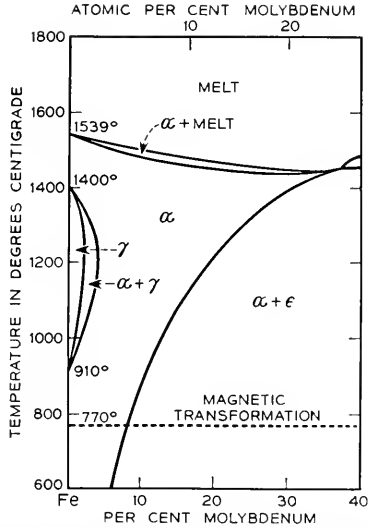

Fig. 23—Phase diagram of iron-rich iron-molybdenum alloys, showing solid solubility curve important in the precipitation-hardening process.

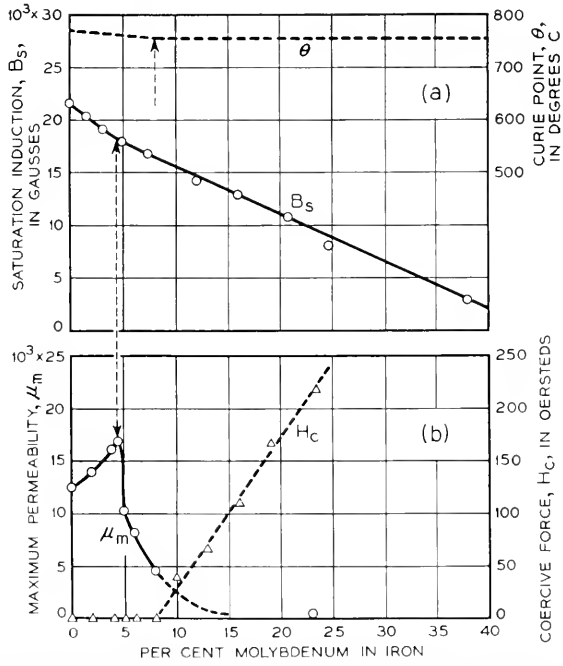

Fig. 24—Change of structure-insensitive properties (θ and B_s) and structure-sensitive properties (μ_m and H_c) with the composition when precipitation-hardening occurs.

When the amount of the second phase is considerable (as in the 15% Mo alloy) it is common practice to quench the alloy from a temperature at which it is a single phase (e.g. 1100 or 1200°C) and so maintain it temporarily as such, and then to heat it to a temperature (e.g., 600°C) at which diffusion proceeds at a more practical rate. During the latter step the second phase separates slowly enough so that it can easily be stopped at the optimum point, after a sufficient amount has been precipitated but before diffusion has been permitted to relieve the strains caused by the precipitation. A conventional heat treatment for precipitation-hardening of this kind, used on many permanent magnet materials, has already been given in Fig. 16.

In some respects the development of atomic order in a structure is like the precipitation of a second phase. When small portions of the material become ordered and neighboring regions are still disordered, severe local strains may be set up in the same way that they are during the precipitation hardening described above. The treatment used to establish high strains is the same as in the more conventional precipitation hardening. The decomposition of an ordered structure in the iron-nickel-aluminum system has been held responsible, by Bradley and Taylor,[10] for the good permanent magnet qualities of these alloys.

Some of the common permanent magnets, heat treated to develop internal strains by precipitation of a second phase, or by the development of atomic ordering, are described in Table IV.

The changes in properties to be expected when the composition varies across a phase boundary of a binary system are shown schematically by the curves of Fig. 25.

Impurities

The principle of precipitation hardening, as just described, applies also to the lowering of permeability by the presence of accidental impurities. For example, the solubilities of carbon, oxygen and nitrogen in iron, described by the curves of Fig. 26, are quite similar in form to the curve separating the α and $\alpha + \epsilon$ areas of the iron-molybdenum system of Fig. 23; the chief difference is that the scale of composition now corresponds to concentrations usually described as impurities. One expects, then, that the presence of more than 0.04 per cent of carbon in iron will cause the permeability of an annealed specimen to be considerably below that of pure iron. The amount of carbon present in solid solution will also affect the magnetic properties.

Because the amounts of material involved are small, it is difficult to carry out well defined experiments on the effects of each impurity, especially in

[10] A. J. Bradley and A. Taylor, *Proc. Roy. Soc.* (London) 166, 353–75 (1938).

TABLE IV

Some Useful Permanent Magnets and Their Properties

Name	Composition (Per cent)*	Fabrication	Heat Treatment	H_c	B_r	Mechanical Properties
Carbon Steel	1Mn, 0.9C	HR, PM	Q800	50	10000	H, S
Tungsten Steel	5W, 0.3Mn, 0.7C	HR, PM	Q850	70	10300	H, S
Chromium Steel	3.5Cr, 0.3Mn, 0.9C	HR, PM	Q830	65	9700	H, S
Cobalt Steel	36Co, 4Cr, 5W, 0.7C	HR, PM	Q950	240	9500	H, S
Remalloy (Comol)	17Mo, 12Co	HR, PM	Q1200, B700	250	10500	H
Alnico 2	12Co, 17Ni, 10Al, 6Cu	C, G	A1200, B600	550	7200	H, B
Alnico 5	24Co, 14Ni, 8Al, 3Cu	C, G	A1300, **B600	550	12500	H, B
Alnico 12	35Co, 18Ni, 6Al, 8Ti	C, G	Cast, B650	950	5800	H, B
Alcomax	25Co, 11Ni, 8Al, 6Cu	C, G	A1300, **B600	550	12500	H, B
Vicalloy	52Co, 10V	C, Cr, PM	B600	300	8800	D
Cunife	20Ni, 60Cu	C, Cr, PM	B600	550	5400	D
Platinum-Cobalt	77Pt, 23Co	C, Cr, PM	Q1200, B650	2600	4500	D
Silmanal	87Ag, 9Mn, 4Al	C, Cr, PM		6000†	550	D

* Remainder iron
Q—quenched from indicated centigrade temperature in oil
A—cooled in air from indicated temperature
B—baked at indicated temperature
HR—hot rolled
CR—cold rolled
PM—punched or machined
C—cast
G—ground
** Cooled in magnetic field
† Coercive force for I = 0
H—hard
B—brittle
D—ductile or malleable
S—strong

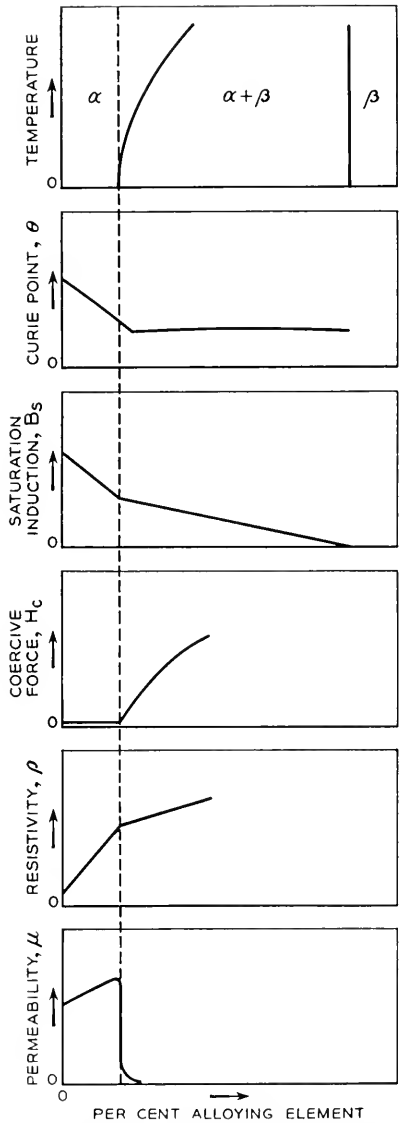

Fig. 25—Diagrams illustrating the changes in various properties that occur when a second phase precipitates.

the absence of disturbing amounts of other impurities. Two examples of the effect of impurities will be given, in addition to Fig. 8. In Fig. 27 Yensen and Ziegler[11] have plotted the hysteresis loss as dependent on carbon content,

[11] T. D. Yensen and N. A. Ziegler, *Trans. Am. Soc. Metals* 24, 337–58 (1936).

the curve giving the mean values of many determinations. The hysteresis decreases rapidly at small carbon contents, when these are of the order of magnitude of the solid solubility at room temperature.

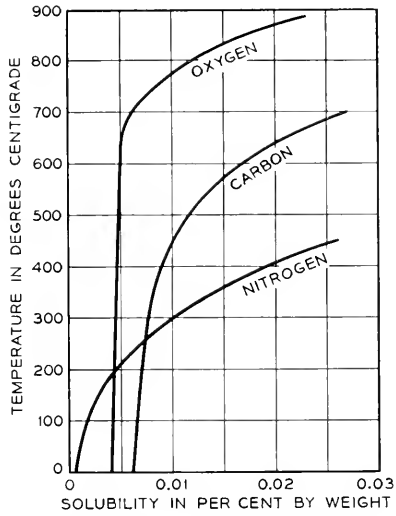

Fig. 26—Approximate solubility curves of carbon, oxygen and nitrogen in iron.

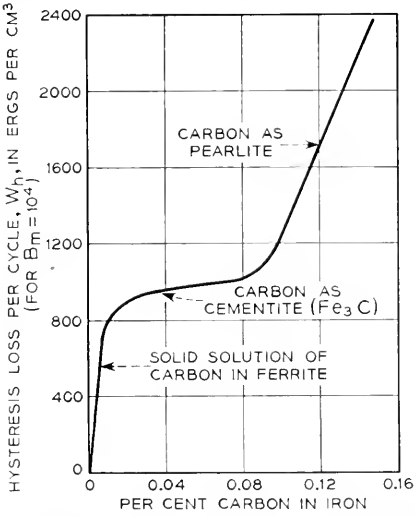

Fig. 27—Effect of carbon content on hysteresis in iron. *Yensen and Ziegler.*

Cioffi[12] has purified iron from carbon, oxygen, nitrogen and sulfur by heating in pure hydrogen at 1475°C, and has measured the permeability

[12] P. P. Cioffi, *Phys. Rev. 39*, 363–7 (1932).

at different stages of purification. Table V shows that impurities of a few thousandths of a per cent are quite effective in depressing the maximum permeability of iron.

Carbon and nitrogen, present as impurities, are known to cause "aging" in iron—that is, the permeability and coercive force of iron containing these elements as impurities will change gradually with time when maintained somewhat above room temperature. As an example, a specimen of iron was maintained for 100 hours first at 100°C, then 150°C, then 100°C, and so on.

TABLE V

Maximum permeability of Armco iron with different degrees of purification, effected by heat treatment in pure hydrogen at 1475°C for the times indicated (P. P. Cioffi).
Analyses from R. F. Mehl (private communication to P. P. Cioffi).

Time of Treatment in Hours	μm	Composition in Per Cent					
		C	S	O	N	Mn	P
0	7000	0.012	0.018	0.030	0.0018	0.030	0.004
1	16000	.005	.010	.003	.0004	—	—
3	30000	.005	.006	.003	.0003	—	—
7	70000	.003	—	.003	.0001	—	—
18	227000	.005	<.003	.003	.0001	.028	.004
Precision of analysis.....		.001	.002	.002	.0001		

The corresponding changes in coercive force are given in the diagram of Fig. 28. A change of about 2-fold is observed.

Some Important Physical Properties

There are many physical characteristics that are important in the study of ferromagnetism from both the practical and the theoretical point of view. These include the resistivity, density, atomic diameter, specific heat, expansion, hardness, elastic limit, plasticity, toughness, mechanical damping, specimen dimensions, and numerous others. In a different category may be mentioned corrosion, homogeneity and porosity. Most of these properties are best discussed in connection with specific materials or properties; only the most important characteristics will be mentioned here. A table of the atomic weights and numbers, densities, melting points, resistivities and coefficients of thermal expansion of the metallic elements, is readily available in the Metals Handbook.

Dissolving a small amount of one element in another increases the *resistivity* of the latter. To show the relative effects of various elements, the common binary alloys of iron and of nickel are shown in Figs. 29 and 30. From a theoretical standpoint it is desirable to understand (1) the relatively

high resistivity of the ferromagnetic elements compared to their neighbors in the periodic table and (2) the relative amounts by which the resistivity of iron (or cobalt or nickel) is raised by a given atomic percentage of various other elements. From a practical standpoint, a high resistivity is usually

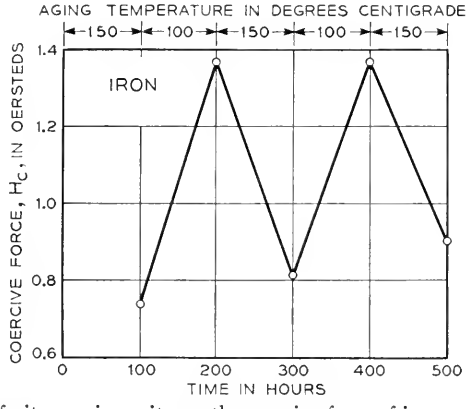

Fig. 28—Effect of nitrogen impurity on the coercive force of iron annealed successively at 100 and 150°C.

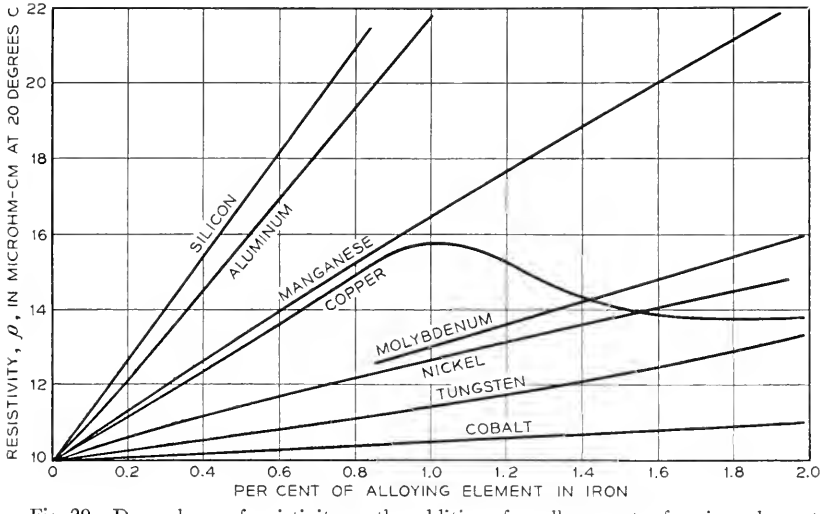

Fig. 29—Dependence of resistivity on the addition of small amounts of various elements to iron.

desirable in order to decrease the eddy-current losses in the material, and so decrease the power wasted and the lag in time between the cause and effect, for example, the time lag of operation of a relay.

Knowledge of the *atomic diameter* is important in considering the effects

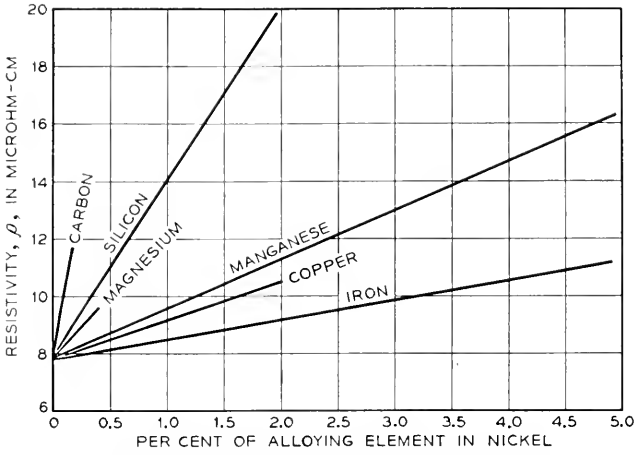

Fig. 30—Resistivity of various alloys of nickel.

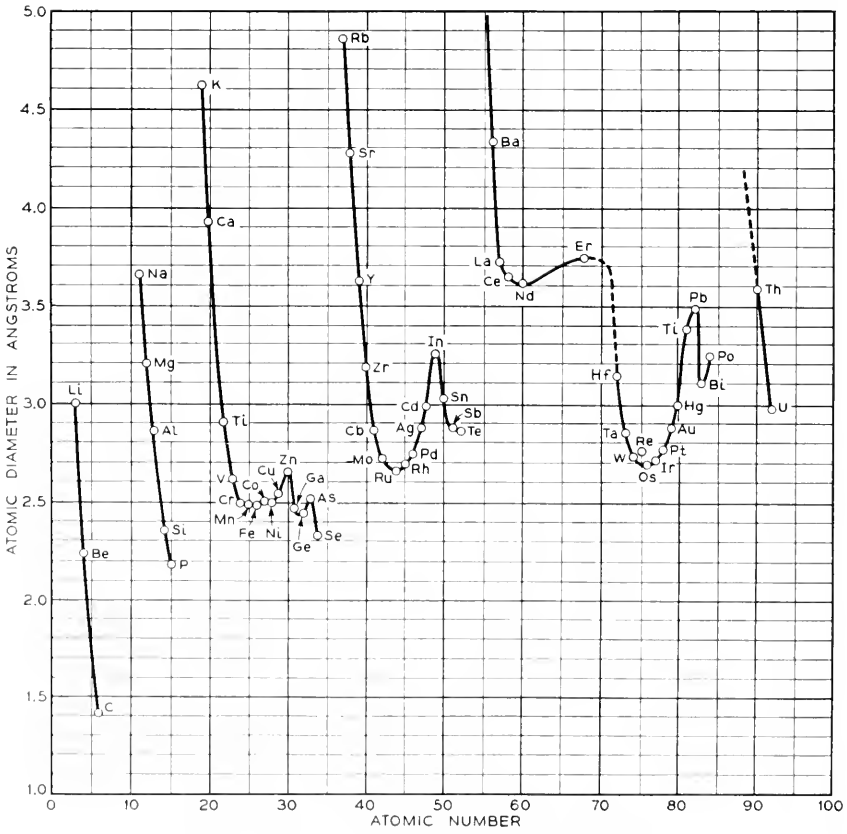

Fig. 31—Atomic diameter of various metallic elements.

of alloying elements, and values for the metallic and borderline elements are shown in Fig. 31. Most of the values are simply the distances of nearest approach of atoms in the element as it exists in the structure stable at room temperature. Atomic diameter is especially important in theory because the very existence of ferromagnetism is dependent in a critical way on the distance between adjacent atoms. This has been discussed more fully in a previous paper.[13]

Even when no phase change occurs in a metal, important *changes in structure* occur during fabrication and heat treatment, and these are complicated and imperfectly understood. When a single crystal is elongated by tension, slip occurs on a limited number of crystal planes that in general are inclined to the axis of tension. As elongation proceeds, the planes on which slip is taking place tend to turn so that they are less inclined to the axis. In this way a definite crystallographic direction approaches parallelism

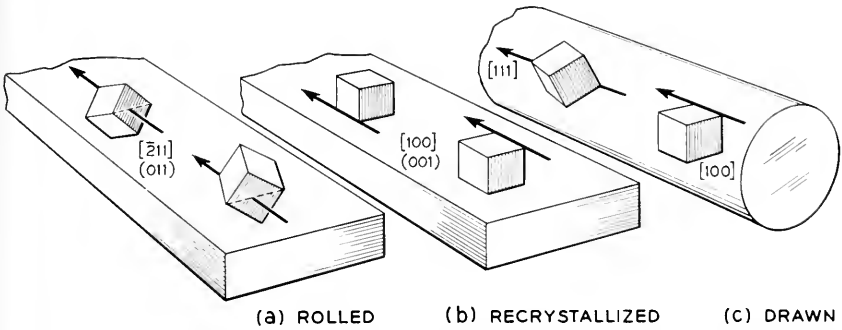

Fig. 32—The preferred orientations of crystals in nickel sheet and wire after fabrication and after recrystallization.

with the length of the specimen. In a similar but more complicated way, any of the usual methods of fabrication cause the many crystals of which it is composed to assume a non-random distribution of orientations, often referred to as preferred or *special orientations*, or *textures*. Some of the textures reported for cold rolled and cold drawn magnetic materials are given in Table VI, taken from the compilation by Barrett.[14] The orientations of the cubes which are the crystallographic units are shown in Fig. 32 (a) and (c) for cold rolled sheets and cold drawn wires of nickel.

Since the magnetic properties of single crystals depend on crystallographic direction (anisotropy), the properties of polycrystalline materials in which there is special orientation will also be direction-dependent. In fact it is difficult to achieve isotropy in any fabricated material, even if fabrication involves no more than solidifying from the melt. The relief of the internal

[13] R. M. Bozorth, *Bell Sys. Tech. Jl.* 19, 1–39 (1940).
[14] C. S. Barrett, Structure of Metals, McGraw Hill, New York (1943).

strains in a fabricated metal by annealing proceeds only slowly at low temperatures (up to 600°C for most ferrous metals) without noticeable grain growth or change in grain orientation, and is designated *recovery*. The principle change is a reduction in the amplitude of internal strains, and this can be followed quantitatively by X-ray measurements. Near the point of complete relief distinct changes occur in both grain size and grain orientation, and the material is said *to recrystallize*. At higher temperatures grain growth increases more rapidly. The specific temperatures necessary for both recovery and recrystallization depend on the amount of previous deformation, as shown in Fig. 33. Special orientations are also present in fabricated materials after recrystallization, and some of these are listed in Table VI and illustrated for nickel in Fig. 32 (b).

As an example of the dependence of various magnetic properties on direction, Fig. 34 gives data of Dahl and Pawlek[15] for a 40 per cent nickel iron

TABLE VI

Preferred Orientations in Drawn Wires and Rolled Sheets, Before and After Recrystallization, and in Castings (Barrett[14])

The rolling plane and rolling direction, or wire axis, or direction of growth, are designated

Metal	Crystal Structure	Drawn wires		Rolled Sheets		As Cast
		As Drawn	Recrystallized	As Rolled	Recrystallized	
Iron.....	BCC	[110]	[110]	(001), [110] and others	(001), 15° to [110]	[100]
Cobalt.....	HCP	—	—	(001)	—	—
Nickel.....	FCC	[111] and [100]	—	(110), [112] and others	(100), [001]	

alloy reduced 98.5 per cent in area by cold rolling and then annealed at 1100°C. After further cold rolling (50 per cent reduction) the properties are as described in Fig. 35.

The mechanical properties ordinarily desirable in practical materials are those which facilitate fabrication. Mild steel is often considered as the nearest approach to an ideal material in this respect. Silicon iron is limited by its brittleness, which becomes of major importance at about 5 per cent silicon; this is shown by the curve of Fig. 36. Permalloy is "tougher" than iron or mild steel and requires more power in rolling and more frequent annealing between passes when cold-rolled, but can be cold-worked to smaller dimensions. If materials have insufficient stiffness or hardness, parts of apparatus made from them must be handled with care to avoid bending and consequent lowering of the permeability. If the hardness is too great the material must be ground to size. This is the case with some permanent magnets.

[15] O. Dahl and F. Pawlek, *Zeits. f. Metallkunde* 28, 230–3 (1936).

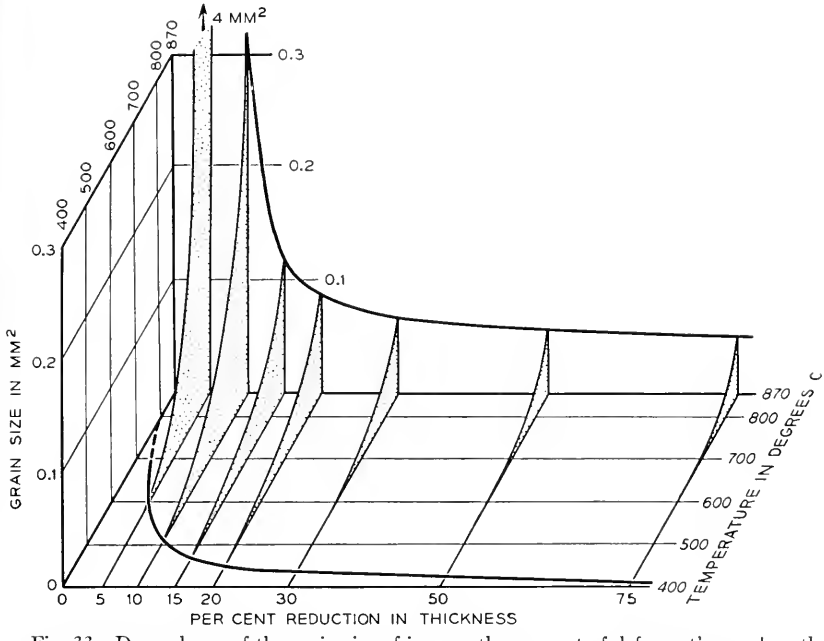

Fig. 33—Dependence of the grain size of iron on the amount of deformation and on the temperature of anneal. *Kenyon.*

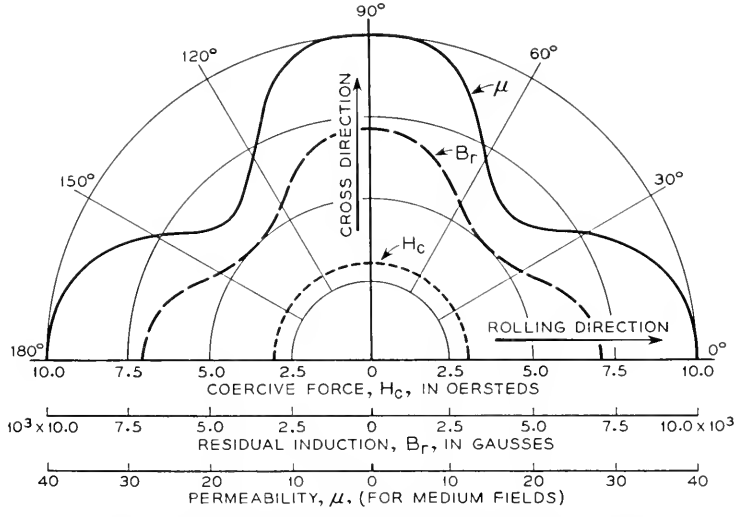

Fig. 34—Variation of magnetic properties with the direction of measurement in a sheet of iron-nickel alloy (40% Ni) severely rolled (98.5%) and annealed at 1100°C.

The effect of size of a magnetic specimen is often of importance. This is well known in the study of *thin films*, and *fine powders* in which the smallest

dimension is about 10^{-4} cm or less. Many studies have been made of thin electrodeposited and evaporated films. Generally it is found that the permeability is low and the coercive force high. The interpretation is uncertain

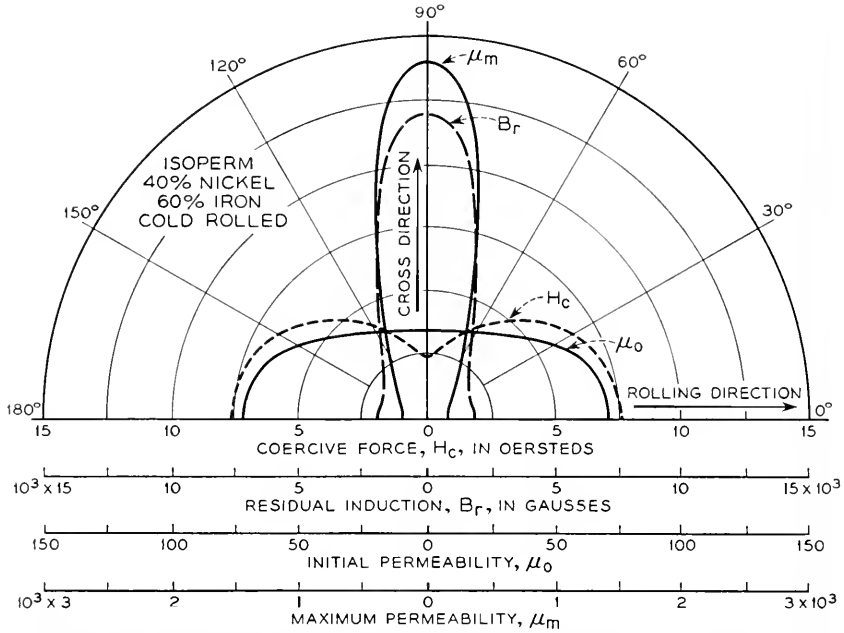

Fig. 35—Properties of the same material as that of Fig. 34, after it has been rolled, annealed, and again rolled.

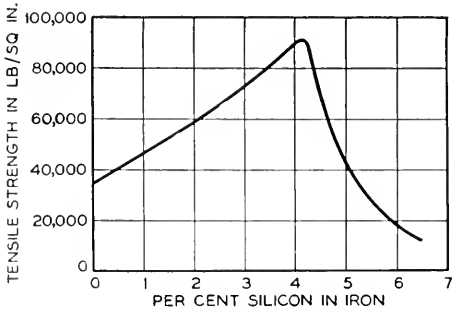

Fig. 36—Variation of the breaking strength of iron-silicon alloys, showing the onset of brittleness near 4 per cent silicon.

because it is difficult to separate the effects of strains and air gaps from the intrinsic effect of thickness, though it is known that each one of these variables has a definite effect. As one example of the many experiments, we

will show here the effect of the thickness of electrodeposited films of cobalt. Magnetization curves are shown in Fig. 37 according to previously unpublished work of the author.

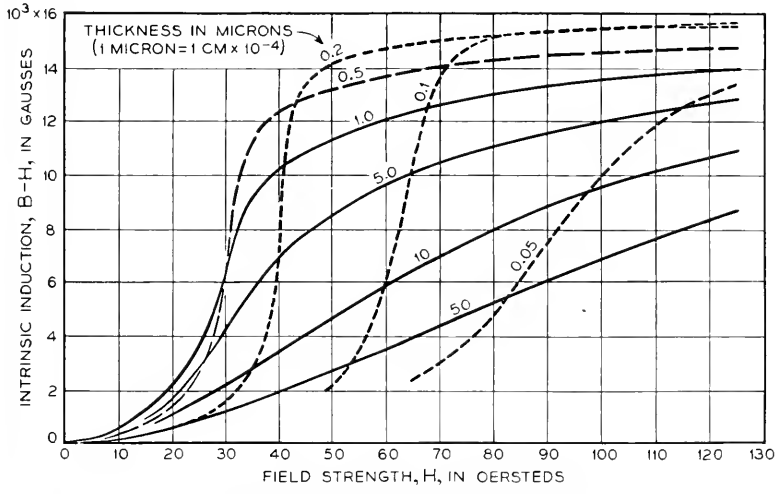

Fig. 37—Dependence of the magnetization curves of pure electrodeposited cobalt films on the thickness.

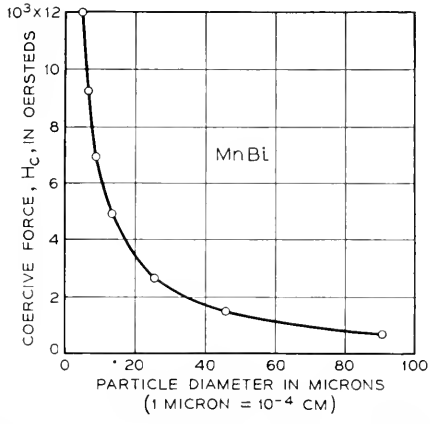

Fig. 38—Dependence of coercive force on the particle size of MnBi powder. *Guillaud.*

The high coercive force obtained in fine powders by Guillaud[16] is one of the most clear cut examples of the intrinsic effect of particle size. The coercive force increases by a factor of 15 as the size decreases to 5×10^{-4} cm (Fig. 38).

[16] C. Guillaud, Thesis, Strasbourg (1943).

Properties Affected by Magnetization

In addition to the magnetization, other properties are changed by the direct application of a magnetic field. Some of these, and the amounts by which they may be changed, are as follows:

Length and volume (magnetostriction) (0.01%)
Electrical resistivity (5%)
Temperature (magnetocaloric effect; heat of hysteresis) (1°C)
Elastic constants (20 per cent)
Rotation of plane of polarization of light (Kerr and Faraday effects) (one degree of arc)

In addition to these properties there are others that change with temperature because the magnetization itself changes. Thus there is "anomalous" temperature-dependence of:

Specific heat
Thermal expansion
Electrical resistivity
Elastic constants
Thermoelectric force

and of other properties below the Curie point of a ferromagnetic material, even when no magnetic field is applied.

Also associated with ferromagnetism are galvanomagnetic, chemical and other effects.

Technical Articles by Bell System Authors Not Appearing in the Bell System Technical Journal

Measurement Method for Picture Tubes. M. W. BALDWIN.[1] *Electronics*, V. 22, pp. 104–105, Nov., 1949.

Diffusion in Binary Alloys.† J. BARDEEN.[1] *Phys. Rev.*, V. 76, pp. 1403–1405, Nov. 1, 1949.

ABSTRACT—Darken has given a phenomenological theory of diffusion in binary alloys based on the assumption that each constituent diffuses independently relative to a fixed reference frame. It is shown that diffusion via vacant lattice sites leads to Darken's equations if it is assumed that the concentration of vacant sites is in thermal equilibrium. Grain boundaries and dislocations may act as sources and sinks for vacant sites and act to maintain equilibrium. The modifications required in the equations if the vacant sites are not in equilibrium are discussed.

Variable Phase-Shift Frequency-Modulated Oscillator. O. E. DE LANGE.[1] *I.R.E., Proc.*, V. 37, pp. 1328–1331, Nov., 1949.

ABSTRACT—The theory of operation of a phase-shift type of oscillator is discussed briefly. This oscillator consists of a broad-band amplifier, the output of which is fed back to the input through an electronic phase-shifting circuit. The instantaneous frequency is controlled by the phase shift through this latter circuit. True FM is obtained in that frequency deviation is directly proportional to the instantaneous amplitude of the modulating signal and substantially independent of modulation frequency.

A practical oscillator using this circuit at 65 mc is described.

Erosion of Electrical Contacts on Make.† L. H. GERMER[1] and F. E. HAWORTH.[1] *Jl. Applied Phys.*, V. 20, pp. 1085–1108, Nov., 1949.

ABSTRACT—When an electric current is established by bringing two electrodes together, they necessarily discharge a capacity. Unless the current which is set up is above 1 ampere, the erosion which is produced in a low voltage circuit is appreciable only when the capacity is of appreciable size and when it is discharged very rapidly by an arc. When the arc occurs, its energy is dissipated almost entirely upon the positive electrode and, when the circuit inductance is sufficiently low, melts out a crater intermediate in volume between the volume of metal which can be melted by the energy

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

and that which can be boiled. Some of the melted metal lands on the negative electrode and, with repetition of the phenomenon, results in a mound of metal transferred from the anode to the cathode. This transfer, which is about 4×10^{-14} cc of metal per erg, is the erosion which occurs on the make of electrical contacts.

The arc voltage is of the order of 15. If the initial circuit potential is more than about 50 volts, there may be more than one arc discharge, successive discharges being in opposite directions and resulting in the transfer of metal in opposite directions—always to the electrode which is negative.

The occurrence of an arc is dependent upon the condition of the electrode surfaces and upon the circuit inductance. For "inactive" surfaces an arc does not occur for inductances greater than about 3 microhenries. Platinum surfaces can be "activated" by various organic vapors, and in the active condition they give arcs even when the circuit inductance is greater than this limiting value by a factor of 10^3.

The Conductivity of Silicon and Germanium as Affected by Chemically Introduced Impurities. G. L. PEARSON.[1] Paper presented at A. I. E. E., Swampscott, Mass., June 20–24, 1949. Included in compilation on semiconductors. *Elec. Engg.*, V. 68, pp. 1047–1056, Dec. 1949.

ABSTRACT—Silicon and germanium are semiconductors whose electrical properties are highly dependent upon the amount of impurities present. For example, the intrinsic conductivity of pure silicon at room temperature is 4×10^{-6} (ohm cm)$^{-1}$ and the addition of one boron atom for each million silicon atoms increases this to 0.8 (ohm cm)$^{-1}$, a factor of 2×10^5.

Although such impurity concentrations are too weak to be detected by standard chemical analysis, the use of radioactive tracers and the Hall effect has made it possible to make quantitative measurements at impurity concentrations as small as one part in 5×10^8.

Silicon and germanium are elements of the fourth group of the periodic table with the same crystal structure as diamonds and they have respectively 5.2×10^{22} and 4.5×10^{22} atoms per cubic centimeter. The addition of impurity elements of the third group such as boron or aluminum gives defect or p-type conductivity. Elements from the fifth group such as phosphorous, antimony or arsenic give excess or n-type conductivity.

The conductivity at room temperature, where it has been shown that each impurity atom contributes one conduction charge, is given by equation (1) where N is the number of solute atoms per cubic centimeter.

$$\sigma = A + BN. \tag{1}$$

[1] B.T.L.

The constants A and B for the various alloys investigated are given in the following table:

Alloy	A	B
Si + B	4×10^{-6}	1.6×10^{-17}
Si + P	4×10^{-6}	4.8×10^{-17}
Ge + Sb	1.7×10^{-2}	4.2×10^{-16}

Equation (1) applies to solute atom concentrations as high as 5×10^{19} per cc. At higher concentrations the mobilities are lowered due to increased impurity scattering so that the computed conduction is higher than the measured.

Microstructures of Silicon Ingots.† W. G. PFANN[1] and J. H. SCAFF.[1] *Metals Trans.*, V. 185 (*Jl. Metals*, V. 1) pp. 389–392, June, 1949.

Increasing Space-Charge Waves.† J. R. PIERCE.[1] *Jl. Applied Phys.*, V. 20, pp. 1060–1066, Nov. 1949.

ABSTRACT—An earlier paper presented equations for increasing waves in the presence of two streams of charged particles having different velocities, and solved the equations assuming the velocity of one group of particles to be zero or small. Numerical solutions giving the rate of increase and the phase velocity of the increasing wave for a wide range of parameters, covering cases of ion oscillation and double-stream amplification, are presented here.

Traveling-Wave Oscilloscope. J. R. PIERCE.[1] *Electronics*, V. 22, pp. 97–99, Nov., 1949.

ABSTRACT—This paper describes a 1,000 volt oscilloscope tube with a traveling-wave deflecting system. The tube is suitable for viewing periodic signals with frequencies up to 500 mc. A signal of 0.037 volt into 75 ohms deflects the spot one spot diameter. A few milliwatts input gives a good pattern, so that the tube can be used without an amplifier. The pattern is viewed through a sixty power microscope.

P-type and N-type Silicon and the Formation of Photovoltaic Barrier in Silicon Ingots.† J. H. SCAFF,[1] H. C. THEURERER[1] and E. E. SCHUMACHER.[1] *Metals Trans.*, V. 185 (*Jl. Metals*, V. 1) pp. 383–388, Jan., 1949.

Longitudinal Noise in Audio Circuits. H. W. AUGUSTADT[1] and W. F. KANNENBERG.[1] *Audio Engg.*, V. 34, pp. 22–24, 45, Jan., 1950.

Transistors. J. A. BECKER.[1] Compilation of three papers presented at A. I. E. E. meeting Swampscott, Mass., June 20–24, 1949. *Elec. Engg.*, V. 69, pp. 58–64, Jan., 1950.

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

Application of Thermistors to Control Networks.† J. H. BOLLMAN[1] and J. G. KREER.[1] *I. R. E., Proc.*, V. 38, pp. 20–26, Jan., 1950.

ABSTRACT—In connection with the application of thermistors to regulating and indicating systems, there have been derived several relations between current, voltage, resistance, and power which determine the electrical behavior of the thermistor from its various thermal and physical constants. The complete differential equation describing the time behavior of a directly heated thermistor has been developed in a form which may be solved by methods appropriate to the problem.

Sensitive Magnetometer for Very Small Areas.† D. M. CHAPIN.[1] *Rev. Sci. Instruments*, V. 20, pp. 945–946, Dec., 1949.

ABSTRACT—A vibrating wire system for measuring weak magnetic fields is described for use in very small spaces. Quartz crystals are used for drivers to get sufficient velocity with very small displacements. To adjust the driving voltage to correspond exactly to the natural crystal frequency, the crystal is also used to regulate the oscillator.

Method of Calculating Hearing Loss for Speech from an Audiogram.† H. FLETCHER.[1] *Acoustical Soc. Am., Jl.*, V. 22, pp. 1–5, Jan., 1950.

ABSTRACT—The question frequently arises, Can one compute the hearing loss of speech from the audiogram and thus make it unnecessary to make a speech test after the hearing loss for several frequencies has been recorded. This paper shows that this can be done by taking a weighted average of the exponentials of the hearing loss at each frequency. Or if β_s is the hearing loss for speech and β_f the hearing loss at each frequency,

$$10^{(\beta_s/10)} = \int G\, 10^{(\beta_f/10)} df$$

The weighting factor G was determined by Fletcher and Galt from threshold measurements of speech coming from filter systems. As specifically applied to the case of hearing loss at the five frequencies 250, 500, 1000, 2000 and 4000 cps, the above equation is approximately equivalent to

$$\beta_s = -10 \log [.01 \times 10^{-(\beta_1/10)} + .13 \times 10^{-(\beta_2/10)} + .40 \times 10^{-(\beta_3/10)} + .38 \times 10^{-(\beta_4/10)} + .08 \times 10^{-(\beta_5/10)}]$$

where β_1 is hearing loss at 250 cps
β_2 is hearing loss at 500 cps
β_3 is hearing loss at 1000 cps
β_4 is hearing loss at 2000 cps
β_5 is hearing loss at 4000 cps

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

Designing for Air Purity. A. M. HANFMANN.[2] *Heating & Ventilating*, V. 47, pp. 59–64, Jan., 1950.

Reciprocity Pressure Response Formula Which Includes the Effect of the Chamber Load on the Motion of the Transducer Diaphragms.† M. S. HAWLEY.[1] *Acoustical Soc. Am., Jl.*, V. 22, pp. 56–58, Jan., 1950.

ABSTRACT—In order to reduce the effects of wave motion in the coupling chamber to permit reciprocity pressure response measurements to higher frequencies, only two of the three transducers involved are coupled at a time to the chamber. Given for these conditions is a derivation of the pressure response formula which includes the effect of the chamber load on the motion of the transducer diaphragms.

Theory of the "Forbidden" (222) Electron Reflection in the Diamond Structure.† R. D. HEIDENREICH.[1] *Phys. Rev.*, V. 77, pp. 271–283, Jan. 15, 1950.

ABSTRACT—The dynamical or wave mechanical theory of electron diffraction is extended to include several diffracted beams. In the Brillouin zone scheme this is equivalent to terminating the incident crystal wave vector at or near a zone edge or corner. The problem is then one of determining the energy levels and wave functions in the neighborhood of a corner. The solution of the Schrödinger equation near a zone corner is a linear combination of Bloch functions in which the wave vectors are determined by the boundary conditions and the requirement that the total energy be fixed. This leads to a multiplicity of wave vectors for each diffracted beam giving rise to interference phenomena and is an essential feature of the dynamical theory.

At a Brillouin zone edge formed by boundaries associated with reciprocal lattice points S and O the orthogonality of the unperturbed wave functions in conjunction with the periodic potential requires that another reciprocal lattice point λ be included in the calculation. The indices of λ must be such that $(\lambda_1\lambda_2\lambda_3) = (s_1s_2s_3) - (g_1g_2g_3)$. The perturbation at the zone edge results in non-zero amplitude coefficients Cg, Cs and Cj for the diffracted waves irrespective of whether or not the structure factor for λ , s or g vanishes. This is the basis of the explanation of the (222) reflection and since it arises through perturbation at a Brillouin zone edge or corner the term "perturbation reflection" is advanced to replace the commonly used "forbidden reflection."

The octahedron formed by the (222) Brillouin zone boundaries exhibits an array of lines due to intersections with other boundaries to form edges. This array of lines is called a "perturbation grid" and the condition for the occurrence of a (222) reflection is simply that the incident wave vector terminate on or near a grid line. Numerical intensity calculations are pre-

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[2] W. E. Co.

sented which show that a strong (222) can be accounted for by the dynamical theory.

An impedance network model is briefly discussed which may aid in qualitative considerations of the dynamical theory for the case of several diffracted waves.

Determination of g-Values in Paramagnetic Organic Compounds by Microwave Resonance. A. N. HOLDEN,[1] C. KITTEL,[1] F. R. MERRITT[1] and W. A. YAGER.[1] Letter to the Editor, *Phys. Rev.*, V. 77, pp. 146–147, Jan. 1, 1950.

Nonlinear Coil Generators of Short Pulses.† L. W. HUSSEY.[1] *I.R.E., Proc.*, V. 38, pp. 40–44, Jan., 1950.

ABSTRACT—Small permalloy coils and circuits have been developed which produce pulses well below a tenth of a microsecond in duration with repetition rates up to a few megacycles.

The construction of these coils is described. Low power circuits are dicussed suitable for different types of drive and different frequency ranges.

Subjective Effects in Binaural Hearing. W. KOENIG.[1] Letter to the Editor, *Acoustical Soc. Am., Jl.*, V. 22, pp. 61–62, Jan., 1950.

ABSTRACT—Experiments with a binaural telephone system disclosed some remarkable properties, notably its ability to "squelch" reverberation and background noises, as compared to a system having only one pickup. No explanation has been found for this subjective effect. It was also discovered that a well-known defect in the directional discrimination of binaural systems was remedied by a mechanical arrangement which rotated the pickup microphones as the listener turned his head.

Corrosion Testing of Buried Cables. T. J. MAITLAND.[3] *Corrosion*, V. 6, pp. 1–8, Jan., 1950.

40AC1 Carrier Telegraph System. A. L. MATTE.[1] *Tel. & Tel. Age*, No. 2, pp. 7–9, Feb., 1950.

Giving New Life to Old Equipment. P. H. MIELE.[3] *Bell Tel. Mag.*, V. 28, pp. 154–163, Autumn, 1949.

Thermionic Emission of Thin Films of Alkaline Earth Oxide Deposited by Evaporation.† G. E. MOORE[1] and H. W. ALLISON.[1] *Phys. Rev.*, V. 77, pp. 246–257, Jan. 15, 1950.

ABSTRACT—Monomolecular films of BaO or SrO were deposited by evaporation on clean tungsten or molybdenum surfaces with precautions to eliminate effects caused by excess metal of the oxide or by heating. Thermionic emissions of the same order of magnitude as from commercial oxide cathodes have been obtained from these systems. The results can be explained qualitatively by considering the adsorbed molecules as oriented dipoles. Although

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[3] A. T. & T.

the results may suggest a possible mechanism for a portion of the emission from thick oxide cathodes, there exist serious obstacles to such thin film phenomena as a complete explanation.

Long Distance Finds the Way. W. H. NUNN.[3] *Bell Tel. Mag.*, V. 28, pp. 137–147, Autumn, 1949.

Private Line Services for the Aviation Industry. H. V. ROUMFORT.[3] *Bell Tel. Mag.*, V. 28, pp. 165–174, Autumn, 1949.

Growing and Processing of Single Crystals of Magnetic Metals.† J. G. WALKER,[1] H. J. WILLIAMS[1] and R. M. BOZORTH.[1] *Rev. Sci. Instruments*, V. 20, pp. 947–950, Dec., 1949.

ABSTRACT—Single crystals of nickel, cobalt and various alloys are grown by slow cooling of the melt. They are oriented by optical means and by X-rays, and ground to the desired shape using the technique described.

A Look Around—and Ahead. L. A. WILSON.[3] *Bell Tel. Mag.*, V. 28, pp. 133–136, Autumn, 1949.

† A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[3] A. T. & T.

Contributors to this Issue

R. M. Bozorth, A.B., Reed College, 1917; U. S. Army, 1917–19; Ph.D. in Physical Chemistry, California Institute of Technology, 1922; Research Fellow in the Institute, 1922–23. Bell Telephone Laboratories, 1923–. As Research Physicist, Dr. Bozorth is engaged in research work in magnetics.

R. W. Hamming, B.S. in Mathematics, University of Chicago, 1937; M.A. in Mathematics, University of Nebraska, 1939; Ph.D. in Mathematics, University of Illinois, 1942. Dr. Hamming became interested in the use of large scale computing machines while at Los Alamos, New Mexico, and has continued in this field since joining the Bell Telephone Laboratories in 1946.

W. P. Mason, B.S. in E.E., University of Kansas, 1921; M.A., Ph.D., Columbia, 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged principally in investigating the properties and applications of piezoelectric crystals and in the study of ultrasonics.

J. R. Pierce, B.S. in Electrical Engineering, California Institute of Technology, 1933; Ph.D., 1936. Bell Telephone Laboratories, 1936–. Dr. Pierce has been engaged in the study of vacuum tubes.

VOLUME XXIX JULY, 1950 NO. 3

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

50¢ per copy

$1.50 per Year

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York 7, N. Y.

Leroy A. Wilson, *President* | Carroll O. Bickelhaupt, *Secretary* | Donald R. Belcher, *Treasurer*

SUBSCRIPTIONS

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

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XXIX *July, 1950* *No. 3*

Principles and Applications of Waveguide Transmission

By GEORGE C. SOUTHWORTH

Under the above title, D. Van Nostrand Company, Inc. will shortly publish the book from which the following article is excerpted. Dr. Southworth is one of the leading authorities on waveguides and was one of the first to foresee the great usefulness that this form of transmission might offer. The editors of the Bell System Technical Journal are grateful for permission to publish here parts of the preface and the historical introduction and chapter 6 in its entirety.

Preface

Though it has been scarcely fifteen years since the waveguide was proposed as a practicable medium of transmission, rather important applications have already been made. The first, which was initiated several years ago, was in connection with radar. A more recent and possibly more important application has been in television where waveguide methods provide a very special kind of radio for relaying program material cross-country from one tower top to another. Already Boston and New York have been connected by this means and shortly Chicago and intervening cities will be added. Other networks extending as far west as the Pacific may be expected. It is reasonable to expect that these two applications will be but the beginning of a more general use.

Interest in the subject of waveguide transmission is not limited to commercial application alone. A comparable interest, perhaps less readily evaluated but nevertheless extremely important, lies in its usefulness in teaching important physical principles. For example there are many concepts that follow from the electromagnetic theory that, in their native mathematical form, may appear rather abstract. However, when translated to phenomena actually observed in waveguides, they become very real indeed. As a result, these new techniques have already assumed a place of considerable importance in the teaching of electrical engineering and applied physics both in lecture demonstrations and in laboratory exercises. It is to be expected

that they will be used even more extensively as their possibilities become better appreciated.

Interest in waveguides has been greatly enhanced by the fact that they brought with them a series of extremely interesting methods of measurement, comparable both in accuracy and scope, with similar measurements previously made only at the lower frequencies. This extension of the range over which electrical measurements may be made has contributed also to neighboring fields of research. One early application led to the discovery of centimeter waves in the sun's spectrum. Another led to important new information about the earth's atmosphere. Still another contributed to the study of absorption bands in gases, particularly bands in the millimeter region. Also of great importance was its contribution to our knowledge of the properties of materials for it led at a fairly early date to measurements at higher frequencies than heretofore of the primary constants, permeability, dielectric constant and conductivity—all for a wide array of substances ranging from the best insulators to the best conductors and including many of the so-called semi-conductors. It is because this new art has already attained considerable stature and is already showing promise as an educational medium that this book has been prepared.

CHAPTER I

INTRODUCTION

1.5 Early History of Waveguides

That it might be possible to transmit electromagnetic waves through hollow metal pipes must have occurred to physicists almost as soon as the nature of electromagnetic waves became fully appreciated. That this might actually be accomplished in practice was probably in considerable doubt, for certain conclusions of the mathematical theory of electricity seemed to indicate that it would not be possible to support inside a hollow conductor the lines of electric force of which waves were assumed to consist. Evidence of this doubt appears in Vol. I (p. 399) of Heaviside's "Electromagnetic Theory" (1893) where, in discussing the case of the coaxial conductor, the statement is made that "it does not seem possible to do without the inner conductor, for when it is taken away we have nothing left on which tubes of displacement can terminate internally, and along which they can run."

Perhaps the first analysis suggesting the possibility of waves in hollow pipes appeared in 1893 in the book "Recent Researches in Electricity and Magnetism" by J. J. Thomson. This book, which was written as a sequel to Maxwell's "Treatise on Electricity and Magnetism," examined mathematically the hypothetical question of what might result if an electric charge

should be released on the interior wall of a closed metal cylinder. This problem is even now of considerable interest in connection with resonance in hollow metal chambers. The following year Joseph Larmor examined as a special case of electrical vibrations in condensing systems the particular waves that might be generated by spark-gap oscillators located in hollow metal cylinders. A more complete analysis relating particularly to propagation through dielectrically-filled pipes both of circular and rectangular cross section was published in 1897 by Lord Rayleigh. Later (1905) Kalähne examined mathematically the possibility of oscillations in "ring-shaped" metal tubes. Still later (1910) Hondros and Debye examined mathematically the more complicated problem of propagation through dielectric wires. Transmission through hollow metal pipes was also considered by Dr. L. Silberstein in 1915.

As regards experimental verification, it is of interest that Sir Oliver Lodge as early as 1894 approached but probably did not quite realize actual waveguide transmission. In a demonstration lecture on electric waves given before the Royal Society, he used, as a source of waves, a spark oscillator mounted inside a "hat-shaped" cylinder. An illustration published later suggests that the length of the cylinder was only slightly greater than its diameter. There is no very definite evidence that the short cylinder functioned as a waveguide or that such a function was discussed in the lecture. Perhaps of greater significance were some experiments reported a year later by Viktor von Lang who used pipes of appreciable length and repeated for electric waves the interference experiment that had been performed for acoustic waves by Quincke some years earlier. Other similar experiments were later performed by Drude and by Weber.

About 1913 Professor Zahn of the University of Kiel became interested in this problem and assigned certain of its aspects to two young candidates for the doctorate, Schriever and Reuter by name. They had barely started when World War I broke out, and both left for the front. Zahn continued this work until he was called a year later. It is reported that by this time he had succeeded in propagating waves through cylinders of dielectric, but it is understood that he did little or no quantitative work. Reuter was killed at Champagne in the autumn of 1915, but Schriever survived and returned to complete his thesis in 1920, using for his source the newly available Barkhausen oscillator.

The contributions of Thomson, Rayleigh, Hondros and Debye, and Silberstein were, of course, purely mathematical. Those of von Lang, Weber, Zahn and Schriever were experimental, but they were of rather limited scope. The concept of the hollow pipe as a useful transmission element, for example as a radiator or as a resonant circuit, apparently did not exist at these early dates. Nothing was yet known quantitatively about attenuation,

and little or nothing of the present-day experimental technique had yet appeared. At this time, the position of this new art was perhaps comparable with that of radio prior to the time of Marconi.

The history of waveguides changed abruptly about 1933 when it was shown that they could be put to practical use. Several patent applications were filed,[1] and numerous scientific papers were published. More recently a great many papers have appeared, too many in fact for detailed consideration at this time. Three of the earlier papers are mentioned in the footnote below.[2] Others will be referred to in the text that follows.

The writer's interest in guided waves stems from some experiments done in 1920 when such waves were encountered as a troublesome spurious effect while working with Lecher wires in a trough of water. In one case there were found, superimposed on the waves that might normally travel along two parallel conductors, other waves having a velocity that somehow depended on the dimensions of the trough. These may now be identified as being the so-called dominant type. In another case, the depth of water was apparently at or near "cut-off," and conditions were such that water waves in the trough gave rise to depths that were momentarily above cut-off, followed a moment later by depths that were below cut-off. This led not only to variations in power at the receiving end of the trough but also to variations in the plate current of the oscillator supplying the wavepower. Indeed these effects could be noted even when the wires were removed from the trough. These waves were recognized as being roughly like those described the same year by Schriever.[3]

Several years later this work was resumed and since that time a continued effort has been made to develop from fundamental principles of waveguide transmission a useful technique for dealing with microwaves. The earliest of these experiments consisted of transmitting electromagnetic waves through tall cylinders of water. Because of the high dielectric constant of water, waves which were a meter long in air were only eleven centimeters long in water. Thus it became possible to set up in the relatively small space of one of these cylinders many of the wave configurations predicted by theory. In addition it was possible, by producing standing waves, to measure their apparent wavelength and thereby calculate their phase velocity. Also by investigating the surface of the water by means of a probe,

[1] Reference is made particularly to U.S. Patents 2,129,711 (filed 3/16/33, 2,129,712 (filed 12/9/33), 2,206,923 (filed 9/12/34) and 2,106,768 (filed 9/25/34).

[2] Carson, Mead and Schelkunoff, "Hyper-frequency Waveguides—Mathematical Theory," *B.S.T.J.*, Vol. 15, pp 310–333, April 1936. G. C. Southworth, "Hyper-frequency Wave Guides—General Considerations and Experimental Results," *B.S.T.J.*, Vol. 15, pp 284–309, April 1936. Also "Some Fundamental Experiments with Waveguides," *Proc. I.R.E.*, Vol. 25, pp 807–822, July 1937. W. L. Barrow, "Transmission of Electromagnetic Waves in Hollow Tubes of Metal," *Proc. I.R.E.*, Vol. 24, pp 1298–1398, October 1936.

[3] The waves actually observed are now known as TE_{10} waves in a rectangular guide, while those described by Schriever are now recognized as TM_{01} waves in a circular guide.

the directions and also the relative intensities of lines of electric force in the wave front could be mapped. It is probable that certain of these modes were observed and identified for the first time.

Shortly afterwards, sources giving wavelengths in air of fifteen centimeters became available and the experimental work was transferred to air-filled copper pipes only 5 inches in diameter. At this time, a 5-inch hollow-pipe transmission line 875 feet in length was built through which both telegraph and telephone signals were transmitted. Measurements showed that the attenuation was relatively small. This early work, which was done prior to January 1, 1934, was described along with other more advanced work in demonstration-lectures and also in papers published in 1936 and 1937.[4]

It was recognized at an early date that a short waveguide line might, with suitable modification, function as a radiator and also as a reactive element. These properties were likewise investigated experimentally, and numerous useful applications were proposed. Descriptions may be found in the numerous patents that followed. These properties were also the subject of several experimental lectures given before the Institute of Radio Engineers and other similar societies by the writer and his associates during the years 1937 to 1939.[5] Included were demonstrations of the waveguide as a transmission line, the electromagnetic horn as a radiator, and the waveguide cavity as a resonator. An adaptation of the waveguide cavity was used to terminate a waveguide line in its characteristic impedance.

From the first, progress was very substantial and by the autumn of 1941 there were known, both from calculation and experiment, the more important facts about the waveguide. In particular, the reactive nature of discontinuities became the subject of considerable study, and impedance matching devices (transformers), microwave filters, and balancers soon followed. Also a wide variety of antennas was devised. Similarly, amplifiers and oscillators as well as the receiving methods followed.

As might be expected, a great many people have contributed in one way or another to the success of this venture. Particular mention should be made of the very important parts played by the author's colleagues, Messrs. A. E. Bowen and A. P. King, who, during its early and less promising period, contributed much toward transforming rather abstract ideas into practical equipment, much of which found important military uses immediately upon the advent of war. Also of importance were the parts played by the author's colleagues, Dr. S. A. Schelkunoff, J. R. Carson, and Mrs. S. P. Meade, who, in the early days of this work, provided a substantial segment of mathematical theory that previously was missing. During the succeeding years, Dr. Schelkunoff, in particular, made invaluable contributions in the form

[4] A description of one of the earlier lectures appears in the Bell Laboratories Record for March 1940. (Vol. XVIII, No. 7, p. 194.)

of analyses which in some cases indicated the direction toward which experiment should proceed and, in others, merely confirmed experiment, while, in still others, gave answers not readily obtainable by experiment alone. In the chapters that follow, the author has drawn freely on Dr. Schelkunoff, particularly as regards methods of analysis.

Beginning sometime prior to 1936, Dr. W. L. Barrow, then of the Massachusetts Institute of Technology, also became interested in this subject and together with numerous associates made very substantial contributions. No less than eight scientific papers were published covering special features of hollow-pipe transmission lines and electromagnetic horns. For several years the work being done at the Massachusetts Institute of Technology and at the Bell Telephone Laboratories probably represented the major portion, if not indeed the only work of this kind in progress, but with the advent of World War II, hundreds or perhaps thousands of others entered the field. For the most part, the latter were workers on various military projects. Starting with the considerable accumulation of unpublished technique that was made freely available to them at the outset of the war, they, along with others in similar positions elsewhere in this country and in Europe, have helped to bring this technique to its present very satisfactory state of development.

CHAPTER VI

A DESCRIPTIVE ACCOUNT OF ELECTRICAL TRANSMISSION

6.0 General Considerations

The preceding four chapters presented the more important steps in the development of the theory of electrical transmission, particularly as it applies to simple networks, wire lines, and waves in free space and in guides. For the most part, the analysis followed conventional methods and made use of the concise and accurate short-hand notation of mathematics. It had for its principal objective the derivation of a series of equations useful in the practical application of waveguides.

Closely associated with the theory of electricity and almost a necessary consequence of it are the numerous concepts and mental pictures by means of which we may explain rather simply the various phenomena observed in electrical practice Though extremely important, this aspect of the theory was not stressed before. Instead it was deferred to the present chapter where it could be considered by itself and from the purely qualitative point of view. It is hoped that this arrangement of material will be of special use to those who find it necessary to substitute for mathematical analysis, simple

models to explain the phenomena which they observe in practice. It is believed that, for these people, this chapter together with a few key formulas taken from the earlier sections will be helpful in gaining a fairly satisfactory understanding of the practical aspects of waveguide transmission.

At the lower frequencies, the current aspect of electricity meets most of the needs and in comparison it is only occasionally that there is a need to discuss lines of electric and magnetic force. In waveguide practice, on the other hand, currents are usually not available for measurement and, although we recognize their reality, they necessarily assume a secondary role. In contrast with currents, we consider the fields present in a waveguide as very real entities and we attach a very great importance to their orientations as well as to their intensities.

6.1 The Nature of Fields of Force

As a suitable introduction to the discussion that follows, we shall review some of the fundamental properties of lines of electric and magnetic force and show pictorially the part that they play in transmission along an ordinary two-wire line.

The Electrostatic Field

As is well known, the concept of the electric field was devised by Faraday to explain the force action between charged bodies. According to his view there exist in the space between the charged bodies, lines or tubes of electric force terminating respectively on positive and negative charges attached to the bodies. These tubes of force are endowed with a tendency to become as short as possible and at the same time to repel, laterally, neighboring lines of force. Their direction at any point is purely arbitrary, but, by subsequent convention, the positive direction is taken from the positively charged body to the negative. This is such that a small positive charge (proton) placed in the field tends to be displaced in the positive sense while an electron tends to move in a negative direction. The force exerted on the unit charge is a measure of the magnitude of the electric intensity **E**. It is measured in volts per meter and, since it has direction as well as magnitude, it is a vector quantity.[1] Figure 6.1-1 illustrates in a general way the arrangement of lines of electrostatic force that are assumed to exist between two oppositely charged spheres. Also shown is a representative vector **E**.

The Magnetostatic Field

In the same way that Faraday provided a satisfactory explanation for the forces between charged bodies, so was he able to explain the forces be-

[1] Black-face type will be used when it seems desirable to emphasize the vector properties of quantities having direction as well as magnitude.

tween magnetized bodies. In the latter case, the two kinds of electrostatic charge are replaced by north-seeking and south-seeking magnetic poles respectively. Similarly the tubes of electric force are replaced by tubes of magnetic force. Roughly speaking, the two kinds of tubes are endowed with analogous properties. Because these magnetic lines are at rest, it is appropriate to speak of them as magnetostatic lines of force and consider them as being comparable but of course not identical with electrostatic lines already discussed. The force exerted on a unit magnetic pole is a measure of magnetic intensity **H**. Like its electric counterpart, it is a vector quantity. In the par-

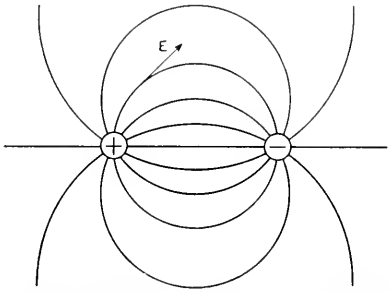

Fig. 6.1-1. Arrangement of lines of electrostatic force in the region between two oppositely charged spheres.

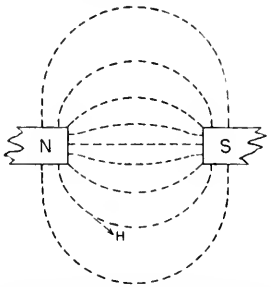

Fig. 6.1-2. Arrangement of lines of magnetostatic force in the region between two oppositely magnetized poles.

ticular system of units used in this text, it is measured in amperes per meter. Figure 6.1-2 illustrates the arrangement of the lines of magnetic force that are assumed to exist between two opposite magnetic poles.

Interrelationship of Electric and Magnetic Fields

As a result of the electromagnetic theory, there are certain properties with which we may endow lines of electric and magnetic force and thereby explain numerous phenomena of electrical transmission. This establishes a relationship between electric and magnetic fields that makes them appear

at times as if they were different aspects of the same thing. They are as follows:

1. *Lines of magnetic force, when displaced laterally, induce in the space immediately adjacent, lines of electric force. The direction of the induced electric force is perpendicular to the direction of motion and also perpendicular to the direction of the original magnetic force. The intensity* **E** *of the induced electric*

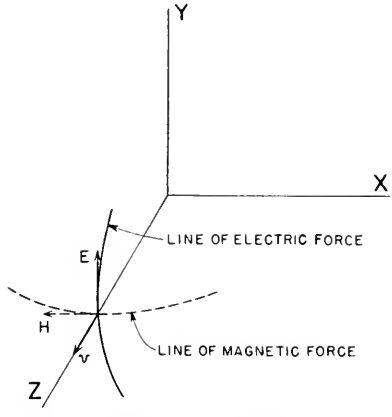

Fig. 6.1-3. Directions of electric vector E and magnetic vector H relative to the velocity v of motion of such lines.

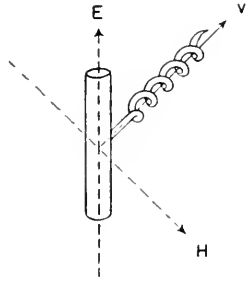

Fig. 6.1-4. Simple corkscrew rule for remembering the directions of E, H and v.

force is proportional to the velocity **v** *of displacement and proportional to the intensity* **H** *of the original lines of magnetic force.*

The directions of the vectors **v**, **E** and **H** are shown in Fig. 6.1-3. They are so related that, when **E** moves clockwise into **H**, it is as though a right-hand screw had progressed in the direction of **v** as shown in Fig. 6.1-4. A convenient short-hand notation used rather generally by mathematicians makes it possible to express these facts by the following vector equation:

$$\mathbf{E} = -\mu(\mathbf{v} \times \mathbf{H}) \qquad (6.1\text{-}1)$$

The quantity μ is the magnetic permeability of the medium under consideration.

2. *Lines of electric force, when displaced laterally, induce in the immediately adjacent space lines of magnetic force. The direction of the induced magnetic force is perpendicular to the direction of motion and also perpendicular to the direction of the original electric force. The intensity* **H** *of the induced magnetic force is proportional to the velocity* **v** *of displacement and proportional to the intensity* **E** *of the original lines of electric force.*

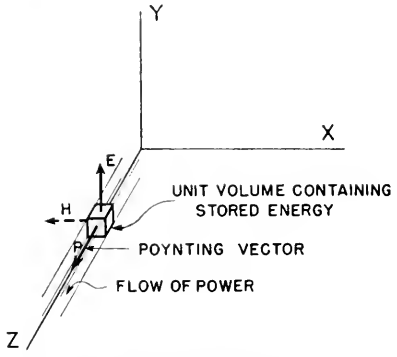

Fig. 6.1-5. Directions of the vectors E and H relative to the Poynting vector P in an advancing wave front.

Again Fig. 6.1-3 and also the right-hand or cork-screw rule apply. In the short-hand notation these facts may be expressed by the following vector equation:

$$\mathbf{H} = \epsilon(\mathbf{v} \times \mathbf{E}) \tag{6.1-2}$$

In this equation, ϵ is the dielectric constant of the medium.[2]

3. *When an electric field of intensity* **E** *is translated laterally, it together with its associated magnetic field* **H** *represents a flow of energy. The direction of the flow of energy is perpendicular to both* **E** *and* **H** *and is therefore in the direction of the velocity* **v**. *The magnitude of the energy flow per unit volume across a unit area measured perpendicular to* **v** *is proportional to the product of the electric intensity* **E** *and the magnetic intensity* **H**. *It may be designated by the vector* **P**.

The relative directions of the vectors **P**, **E**, and **H** are shown in Fig. 6.1-5. The energy per unit volume moves with a velocity expressed by

$$v = \frac{1}{\sqrt{\mu\epsilon}} \tag{6.1-3}$$

[2] The values of permeability μ and dielectric constant ϵ appearing in these equations are not the values found in most tables of the properties of materials. As here given μ is smaller than the usual value μ_r by a factor of 1.257×10^{-6} while ϵ is smaller than ϵ_r by a factor of 8.854×10^{-12}. The use of these special values leads to certain mathematical simplifications.

It therefore corresponds to a flow of power. In the notation just referred to, it may be expressed by the vector equation

$$\mathbf{P} = \mathbf{E} \times \mathbf{H} \tag{6.1-4}$$

4. *Lines of force exhibit the properties of inertia. They therefore resist acceleration.*

Other principles not quite so fundamental but nevertheless useful in application are:

5. *Lines of force are under tension and at the same time are under lateral pressure.*

6. *For perfect conductors there can be no tangential component of electric force.* That is to say, lines of electric force when attaching themselves to a perfect conductor must approach perpendicularly. This is substantially true also for common metals such as copper.

In passing it is well to point out that the first principle is really that by which the ordinary dynamo operates. The second is, for practical purposes, Oersted's Principle, if we assume that the lines of electric force are attached to charges flowing in near-by conductors. The third is known as the Poynting Principle. It has a wide field of application contributing very materially to the physical pictures of both radio and waveguide transmission. When applied to the very simple case of low frequencies propagated along a transmission line, it gives a result that is in keeping with the usual view that the power transmitted is equal to the product of the total voltage times the total current. The fourth principle is useful in explaining qualitatively how radiation from an antenna takes place. The usefulness of these four principles will be made more evident by the examples that follow.

6.2 Transmission of Power along a Wire Line

Direct Current

According to the Poynting concept, one may think of an ordinary dry cell as two conductors combined with chemical means for producing a continuous supply of lines of electric force. This need not be counter to the accepted views concerning electrolysis, for we may think of these lines of force as being attached to ionic charges incidental to dissociation. As long as the cell is on open circuit, these lines of electric force remain in a static condition in which many are grouped in the neighborhood of the terminals of the cell as shown in Fig. 6.2-1(a). In this state of equilibrium, the forces of lateral pressure are balanced by the forces of tension. There is no motion and hence no flow of power. For an ordinary dry cell such as used in flashlights, the electric intensity $\mathbf{E}$ will depend on the spacing of electrodes, but it may be as much as 200 volts per meter If we attach to the dry cell two parallel wires spaced perhaps a centimeter apart with their remote ends open, electro-

static lines will be communicated to the wires, thereby providing a distribution roughly like that shown in Fig. 6.2-1(b). Except at the moment of contact, there is no motion of the lines of electric force and therefore no magnetic field and, accordingly, there can be no flow of power. The final configuration is to be regarded as the resultant of the forces of tension and lateral pressure. The electric intensity, **E**, measured in volts per meter at any point along the line, may be altered at will, merely by changing the spacing.

If, next, we close the remote end of the line by substituting a conducting wire for the particular line of force shown as a heavy line in Fig. 6.2-1(c), the adjacent lines of electric force will collapse on the terminating conductor,

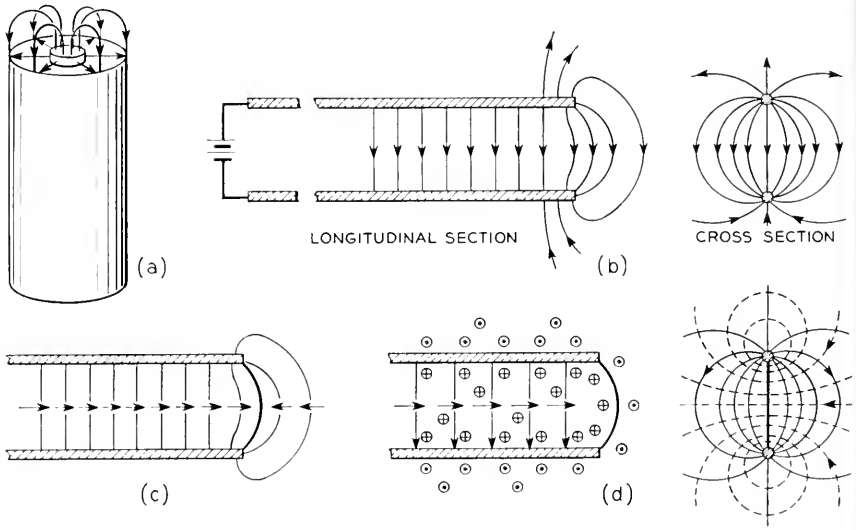

Fig. 6.2-1. Lines-of-force concept applied to the transmission of d-c power along a wire line.

as opposing charges unite. This removes the lateral pressure on the neighboring lines with the result that the whole assemblage starts moving forward. Each line of force meets in its turn the fate of its forerunners, thereby delivering up its energy to the resistance as heat. As soon as the lateral pressure at the cell is relieved, chemical equilibrium is momentarily destroyed and more lines of force are manufactured to fill the gaps of those that have gone before. All of this is, of course, at the expense of chemical action.

According to the electromagnetic theory, as set forth in the second principle, this is but a part of the story of transmission. We must add that the motion of the lines of electric force from the dry cell toward the resistance gives rise in the surrounding space to lines of magnetic force in accordance

with Equation 6.1–2 and furthermore the two fields together give rise to component Poynting vectors representing power flow. Each component vector has a magnitude at any point equal to the product of the electric and magnetic intensities there prevailing and a direction at right angles to the two component forces in accordance with Equation 6.1-3. This is illustrated in Fig. 6.2-1(d).

Since the fields reside largely outside the conductors, we conclude that the principal component of power flow is through the space between the wires and not through the wires themselves. If, in the case cited above, there is appreciable resistance in the connecting wires, then we may expect that there will be a small component of energy flowing into the wires to be dissipated as heat. To account for this, we may picture lines of electric force

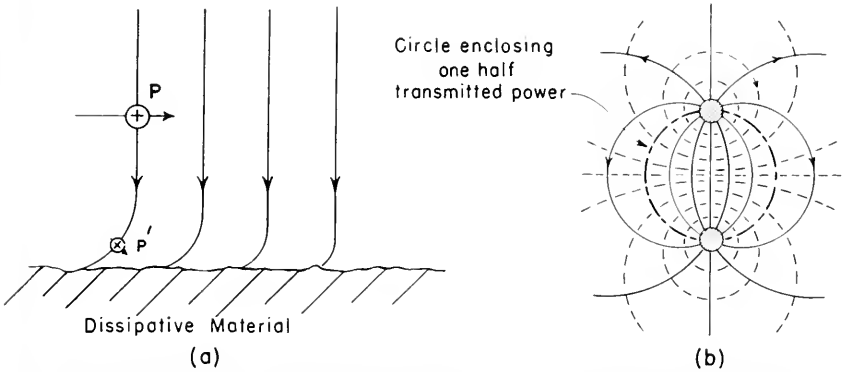

Fig. 6.2-2. Fields of electric and magnetic force and also direction of power flow in the vicinity of conductors. (a) Magnified view showing power flow along a single dissipative wire. (b) Cross-sectional view of parallel-wire line.

which in the immediate vicinity of the conducting wire lag somewhat behind the portions more remote. This is illustrated by Fig. 6.2-2(a) which shows a highly idealized and greatly enlarged section of the field in the immediate vicinity of one of the two dissipative conductors. The very small component of power flowing into the conductor is designated as the vector P' to distinguish it from the much greater power P which we shall assume is being propagated parallel to the conductor.[3]

The magnetic field associated with two cylindrical conductors consists of circles with centers on the line joining the two conductors, whereas the electric field consists of another series of circles orthogonally related to the

[3] For all metals from which conducting lines are ordinarily made, the component of power flowing into the conductor is extremely small compared with the power flowing parallel to its surface. In Fig. 6.2-2(a) therefore, we should regard vector P' as greatly exaggerated in magnitude relative to that of vector P.

first, and having centers on a line at right angles to the first as shown in Fig. 6.2-2(b). The total flow of power through any plane set up perpendicular to the wires is found by adding up the various component products of **E** and **H** from the boundaries of the wires to infinity. The method by which this is carried out is outside of the scope of this chapter, but, as already pointed out, it leads to the same result as obtained by multiplying together the total voltage and the total current. There are two results of this integration that are of special interest. (1) In the case of two parallel cylinders, one-half of the total power flows through the space enclosed by a circle drawn about the wire spacing as a diameter [see Fig. 6.2-2(b)]. The remaining half extends from this circle on out to infinity. (2) Since both the electric and magnetic intensities are greatest in the neighborhood of the wire, most of the total power flow takes place in the immediate vicinity of the wire.

Transmission of A-c Power

If the simple d-c source mentioned previously is replaced by an alternating electromotive force, a variety of phenomena may take place, the more important of which will depend on the frequency of alternation. If this frequency is low (very long wavelength), the line may be relatively short compared with the wavelength, with the result that changes occurring at the source may appear very soon at the remote end. For this case, the observed phenomena will vary sinusoidally with time everywhere along the line, in substantially the same phase. This is the typical alternating-current power line problem[4] and, except for minor details, which we shall not discuss at this time, it does not differ materially from the simple d-c case already covered.

If, on the other hand, the frequency is high (short wavelength), the line may be regarded as being *electrically long*, with the result that sinusoidal changes occurring at the source may not have traveled very far before the direction of flow at the source has changed. The over-all result in extreme cases may become very complicated indeed; for, wavepower may not only be reflected from the remote end of the line but, if there are sharp bends in the line or abrupt changes in spacing, it may be reflected from these points also. The phenomenon observed is usually referred to as *wave interference* and it often leads to *standing waves*. Though described above as complicated, there are many cases where the results of wave interference may be sufficiently simple to be readily visualized. Practical difficulties of various kinds may arise from these effects, but they may also serve very useful purposes. In fact, a substantial portion of our microwave technique is based on wave

[4] The wavelength corresponding to a frequency of 60 cycles per second is five million meters. A commercial power line having a length as great as 100 miles is therefore but 0.03 wavelength long. It is said to be *electrically short*.

interference. Certain specific examples will be discussed later, but first we shall discuss a somewhat simpler case.

The Infinite Line

Let us take, for discussion, a uniform two-wire line that is infinitely long. Waves launched on such a line are assumed to be propagated to infinity. There are no reflected components and hence no wave interference. If the frequency is very high, the forerunners of the lines of force sent out by the source will not have traveled very far when the emf at the source will have reversed its direction. This gives rise at the source to a second group of lines

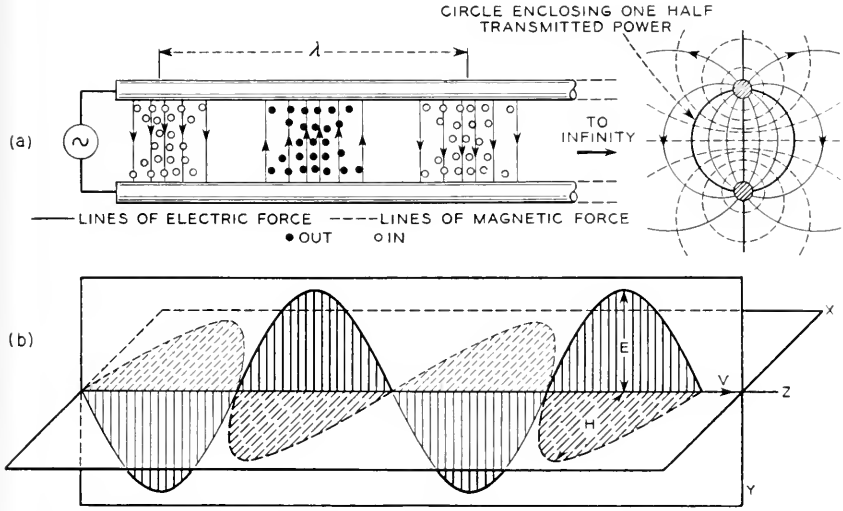

Fig. 6.2-3. (a) Arrangement of lines of electric and magnetic force in both the longitudinal and transverse sections of an infinitely long transmission line. (b) Space relationship between electric vector E and magnetic vector H as observed in a plane containing the two conductors.

of force exactly like the first except oppositely directed. This, in turn, will be followed by a third group identical with the first and a fourth identical with the second and so forth until equilibrium is reached. Because the lines of electric force are in motion, we must expect them to be accompanied by lines of magnetic force. Both are of equal importance. Therefore it is not correct to refer to either alone as a distinguishing feature of the wave. Both components are shown in cross section at the right in Fig. 6.2-3(a).

The distance between successive points of the same electrical phase in a wave is known as the wavelength λ. It depends on the frequency f of alternation and the velocity of propagation v; $\lambda = v/f$. The velocity of propagation in turn depends on the nature of the medium between the two wires. For

air, the velocity v_a is substantially 300,000,000 meters per second (186,000 mi per sec). For other media $v = v_a/\sqrt{\mu_r \epsilon_r}$. Thus it will be seen that, by replacing the air normally found between the two wires of a transmission line by another medium such as oil ($\epsilon_r = 2$ and $\mu_r = 1$), the wavelength will be reduced by a factor of $1/\sqrt{2}$.

If A_0 is the maximum amplitude reached by the oscillating source during any cycle, the amplitude at any time t, measured from an arbitrary beginning, may be expressed by the equation

$$A = A_0 \sin(\omega t + \phi) = A_0 \sin\left(\frac{2\pi}{\lambda} vt + \phi\right) \tag{6.2-1}$$

where ϕ is the initial phase of the amplitude relative to an arbitrary reference angle

If the transmission line is free from dissipation and we choose a datum point in a plane at right angles to the direction of propagation and at a distance far enough from the source that the lines of force have had an opportunity to conform to the wire arrangement and if we designate the electric intensity at this point as E_0 and the corresponding magnetic intensity as H_0, then the electric and magnetic intensities at other corresponding points at a distance z further along the line may be represented by

$$E = E_0 \sin \frac{2\pi}{\lambda} (z - vt)$$

and

$$H = H_0 \sin \frac{2\pi}{\lambda} (z - vt) \tag{6.2-2}$$

These equations are the trigonometric representations of an unattenuated sinusoidal wave of electric intensity and magnetic intensity traveling in a positive direction along the z axis. They are plotted in the yz and xz planes of Fig. 6.2-3(b). An electromagnetic configuration similar to the above but traveling in the opposite direction is given by

$$E = E_0 \sin \frac{2\pi}{\lambda} (z + vt)$$

and

$$H = H_0 \sin \frac{2\pi}{\lambda} (z + vt) \tag{6.2-3}$$

These equations may be further confirmed by plotting arbitrary values on rectangular-coordinate paper. In an infinite line the magnetic intensity **H** and the electric intensity **E** are in the same phase as shown in Fig. 6.2-3.

If the wave is subject to an attenuation of α units per unit distance, possibly due to resistance in the wires, the corresponding components of **E** and **H** are equally attenuated. Either component may be expressed by an equation of the type

$$E = E_0 e^{-\alpha z} \sin \frac{2\pi}{\lambda} (z - vt) \tag{6.2-4}$$

This is a very special form of certain equations appearing in Sections 3.2 and 3.3.

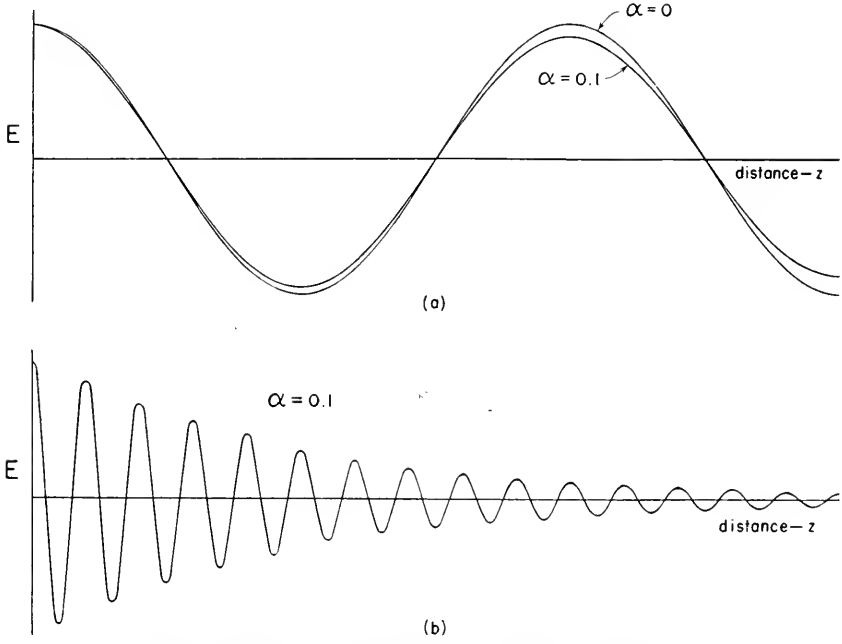

Fig. 6.2-4. Effect of attenuation on an advancing wave front.

If the attenuation is negligible, then $\alpha = 0$ and the term $e^{-\alpha z}$ will be unity. Equation 6.2-4 will then reduce to 6.2-2. If, on the other hand, the attenuation is considerable, the product of α times z will increase rapidly with distance, and the factor $e^{-\alpha z}$ will have the effect of reducing the electric intensity E prevailing at various points along the line. Figure 6.2-4(a) illustrates the variation, with distance, of the electric intensity E for an unattenuated wave $\alpha = 0$. There is included for comparison purposes the case, $\alpha = 0.1$. Figure 6.2-4(b) shows the effect of this rate of attenuation on waves that have traveled for some distance. It is significant that moderate amounts of attenuation have little or no effect on wavelength.

At low frequencies, conductor loss is often the principal cause of attenuation. At high frequency, this loss may be still more important[5] and in addition there may be losses in the medium around the two conductors. The latter is particularly true when the conductors are supported on insulators or are embedded in insulating material. There may also be losses due to lines of force that detach themselves from the wires and float off into the surrounding space (radiation). All three lead to attenuation and may be expressed in terms of an equivalent resistance. They are amenable to calculation for certain special cases.

According to one view of electricity, the individual charges to which lines of force attach themselves are unable to flow through the conductor with the velocity of light. If this is true, lines of force snap along from one charge to the next in a rather mysterious fashion which we will not attempt to picture at this time. This view, like others mentioned previously, tends to relegate the charges and hence the currents to a secondary position.

Although infinitely long transmission lines cannot be constructed in practice, it is possible, by a variety of methods, to approximate this result. In general, a resistance connected across the open end of a short transmission line, of the kind here assumed, absorbs a portion of the arriving wavepower and reflects the remainder. If the resistance is either very large or very small, the reflected power may be very substantial but, by a suitable choice of intermediate values of resistance, the reflected part may be made very small indeed. In the ideal case, the arriving wavepower is completely absorbed. A line connected to this particular value of resistance appears to a generator at the sending end as though it were infinitely long. The particular resistance that can replace an infinite line at any point, without causing reflections, is known as the *characteristic impedance* of the line. This quantity depends on the dimensions and spacings of the two conductors as well as the nature of the medium between. A parallel-wire line, in air, usually has a characteristic impedance of several hundred ohms. A coaxial line filled with rubber often has a characteristic impedance of a few tens of ohms. A line having characteristic impedance connected at its receiving end is said to be *match-terminated*.

Reflections on Transmission Lines

If the transmission line ends in a termination other than characteristic impedance, or if there are discontinuities, due to impedances connected either in series or in shunt with the line, reflections of various kinds will occur.[6] Much of the practical side of microwaves has to do with these reflections.

[5] The losses in most conductors increase with the square root of the frequency.

[6] At the higher frequencies, reflections may also occur at points where the wire spacing changes abruptly. In some instances abrupt changes in wire diameter may be sufficient to cause reflection. These discontinuities may be regarded as changes in characteristic impedance.

A particularly simple form of reflection occurs when the high-frequency transmission line is terminated in a transverse sheet of metal of good conductivity, as for example, copper. An arrangement of this kind is shown in Fig. 6.2-5. As it is difficult to represent a wave front moving toward the reflecting plate, we shall substitute an imaginary thin slice or section of the electromagnetic configuration. A slice of this kind is shown in Fig. 6.2-5(a).

Experiment shows that, at the boundary of the nearly perfect reflector, the transverse electric force E is extremely small. This is consistent with the sixth principle set forth in the previous section which states that there can be no tangential component of electric force at the boundary of a perfect conductor. The result actually observed can be accounted for if it is assumed that the reflecting conductor merely reverses the direction of lines of electric force as they become incident, thereby giving rise to two sets of

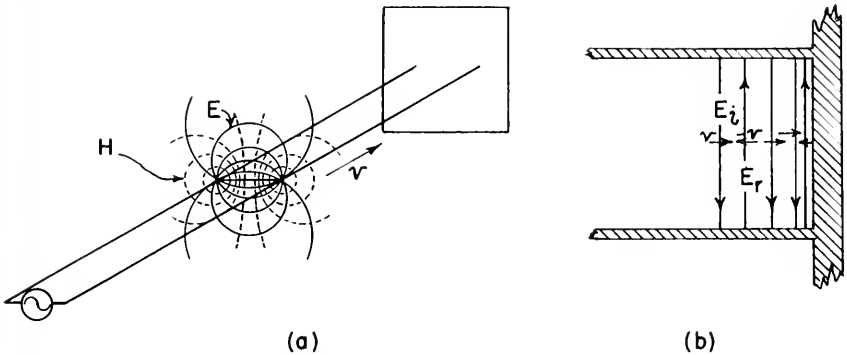

Fig. 6.2-5. (a) Propagation of an electromagnetic wave along a two-wire line terminated by a large conducting plate. (b) Representative lines of force reflected by the conducting plate.

lines of force as shown in Fig. 6.2-5(b), one of intensity $E_i = E$ directed downward in the figure and moving laterally toward the metal sheet (incident wave) and the other of intensity $E_r = -E$ directed upward and moving away from the metal sheet (reflected wave). Accordingly the resultant electric intensity at the surface is zero.

If the reflector is non-magnetic, the magnetic intensity $\mathbf{H}$ will be unaffected by the reflecting material. We find by applying the right-hand rule of Fig. 6.1-4 that the electric intensity $\mathbf{E}_r = -\mathbf{E}$ when combined with $\mathbf{H}$ constitutes a wave that must travel in a negative direction of $\mathbf{v}$. This wave may be represented by Equation 6.2-3. In a similar way the Poynting vector which before reflection is represented by $\mathbf{P} = \mathbf{E} \times \mathbf{H}$ now takes the form $\mathbf{P} = (-\mathbf{E} \times \mathbf{H})$. The negative sign according to the right-hand rule of Fig. 6.1-4 shows that the power approaching the conductor is reflected back upon itself. If $\mathbf{E}$ and $\mathbf{H}$ are respectively equal in magnitude before and after

incidence, the reflection is perfect, and the coefficient of reflection is said to be unity. Bearing in mind that $\mathbf{H}_i = \epsilon(\mathbf{v} \times \mathbf{E})$ before reflection and $\mathbf{H}_r = \epsilon(-\mathbf{v} \times -\mathbf{E})$ after reflection, it is evident that the direction of the magnetic intensity has been unchanged by the process of reflection and that the resultant magnitude at the surface of the metal is $|H_i| + |H_r| = 2|H|$. Thus we see that, at the moment of reflection from a metallic surface, the resultant electric force vanishes and the resultant magnetic force is doubled.

The reflection of waves at the end of the line naturally gives rise to two oppositely directed wave trains. This is a well-known condition for standing waves. Though a complete discussion of standing waves calls for the mathematical steps taken in Section 3.6, there are certain qualitative results that may be deduced from relatively simple reasoning. Some of these deductions will be made in the paragraphs that follow.

If an observer, endowed with a special kind of vision for individual lines of force, were to be stationed at various points along a lossless transmission line as shown in Fig. 6.2-5, he would observe a variety of phenomena as follows. Near the reflector he would observe a waxing and waning of lines of force, both electric and magnetic, corresponding to the arrival of crests and hollows of waves. Also he would observe a similar waxing and waning corresponding to waves leaving the reflector. The sum of the two waves would give rise at the conducting barrier to a resultant electric intensity of zero and to a corresponding magnetic intensity that would oscillate between limits of plus or minus $2H$. Since it is the magnetic component that is the the more evident near the barrier, this region would appear to the observer much like the interior of a coil carrying alternating current.

If the observer were to pass along the line to a point one-eighth wavelength to the left of the reflector, the distance up to the reflector and back would then be a quarter wave and he would then find that at the moment that a wave crest (maximum intensity) was passing on its way toward the reflector a point on the wave corresponding to zero intensity would be returning from the reflector. Adding the corresponding electric and magnetic intensities at this point, he would observe that the electric intensity would not always be zero but instead it would oscillate between limits of plus or minus $\sqrt{2}\,E$. Similarly the corresponding magnetic intensity would no longer oscillate between limits of plus or minus $2H$, but instead it would never reach limits greater than plus or minus $\sqrt{2}\,H$. Thus at this point the electric and magnetic components would have the same average intensity.

If the observer were to move farther along the line, stopping this time at a distance of one-fourth wavelength to the left of the metal plate, the total electrical distance to the barrier and back again would be a half wavelength and he would now find that at the time a crest passed on its way toward the reflector a hollow (maximum negative intensity) would be pass-

ing on its return journey. This time, the resultant electric intensity would oscillate between limits of plus or minus $2E$, and the resultant magnetic intensity would be zero at all times. To this observer then, this quarter-wave point on the line would have many of the characteristics of the interior of a condenser charged by an alternating voltage.

If our observer were to move another one-eighth wave farther along the line, he would note that the resultant electric and magnetic forces would again be equal. Proceeding on to a point one-half wavelength from the metal reflector, he would observe that, at the time crests (maximum positive intensity) were passing on their way toward the reflector, hollows would be returning, and accordingly upon examining the resultant electric intensity he would find it to be zero at all times, whereas the corresponding magnetic intensity would be oscillating between limits of plus or minus $2H$. At this point along the line, he would be unable to distinguish his electrical environment from that prevailing at the metal boundary. The half-wave line, therefore, has had the effect of translating the metal barrier to another point in space a half wave removed.

If the observer were to continue still farther along the line, he would pass, alternately, points where the resultant electric force is zero and other points where the resultant magnetic force is zero. It is important to note that at points in a standing wave where the magnetic force is a maximum, the electric force is a minimum and at points where the electric force is a maximum, the corresponding magnetic force is a minimum. It is customary to call the points of minimum E (or H) "mins," though the term *node* is sometimes substituted. Points of maximum E (or H) are known as "maxs" with the term *loop* as its alternative. If the observer were to measure current and voltage along the line, he would find that points of maximum voltage correspond to maximum E and that points of maximum current correspond to maximum H.

An examination of the energy associated with the incident and reflected waves shows that, except for minor losses not to be considered here, there is as much energy led away from the reflector as is led up to the reflector, and that there is associated with the standing wave a stored or resident energy. The regular arrangement of nodes and loops along a standing wave with minima at half-wave intervals is a very important characteristic, for such points may be located very accurately experimentally, and accordingly wavelength may be measured with considerable precision.

If, instead of terminating the wire line in a large conducting plane assumed previously, it is terminated in a relatively thin cross bar as shown in Fig. 6.2-6, the reflection will assume a somewhat more complicated form. First of all, the thin cross bar will intercept, initially at least, only a portion of the total wave front. The particular lines of force arriving along a plane

containing the two wires will be the first to be reflected and they will behave at reflection much like those already discussed, whereas those outside the plane of the two wires will not be intercepted initially by the thin cross bar but instead will advance for a short distance beyond the end of the line before their forces of tension bring them to rest. These outlying lines of force are represented by the lines designated as c in Fig. 6.2-6. After the first lines of force have been reflected, lateral pressure will be removed from those adjacent, with the result that they will close in and collapse on the conductor at a slightly later time than their neighbors. One over-all result of this process is to make the effective length of such a line slightly greater than the true length. Effects of this kind are observed in practice and they are referred to as *fringing*. Discrepancies between the wavelength as measured in the last section of line where fringing may take place and that measured between other minima along the same line are usually small but

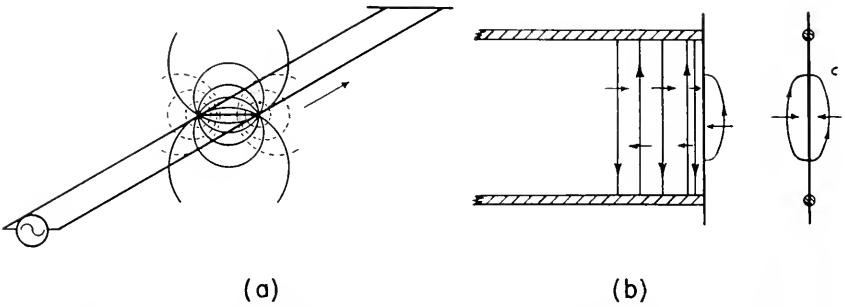

Fig. 6.2-6. (a) Representative transmission line terminated by a conductor of finite dimensions. (b) Nature of reflection by a finite conductor.

they are nevertheless measurable It is also true that, as the wave front approaches a limited barrier of this kind, some of its energy continues on into the space beyond and is lost as radiation. In general, the smaller the barrier, the larger will be the losses.

Consider next a line open at its remote end, as shown in Fig. 6.2-7 In this case, none of the lines of force of the advancing wave is intercepted by a conductor, with the result that a very considerable number momentarily congregate near the end of the line and, because of inertia, they extend into the space beyond as suggested by Fig. 6.2-7(b). This process continues until forces of tension in the lines, still clinging fast to the ends of the wires, bring the assemblage temporarily to rest. At this moment, there is no magnetic component; for v, in the relation $\mathbf{H} = \epsilon(\mathbf{v} \times \mathbf{E})$, is zero while the corresponding electric intensity is approximately $2E$. The lines of electric force, being momentarily at rest, represent energy stored in the electric form.

This static situation is extremely temporary, for the tension momentarily created in the lines of electric force soon forces the configuration as a whole to move backward. As the wave front gets under way, the magnetic force H increases in magnitude in accordance with the relation $\mathbf{H} = \epsilon(\mathbf{v} \times \mathbf{E})$.

The fact that the wave front extends momentarily for a short distance beyond the physical end of the line and requires time to come to rest and get into motion in the reverse direction implies inertia or momentum in the wave front. This is the inertia referred to in the fourth principle mentioned in Section 6.1. In this form of reflection, fringing is usually very evident, and because of fringing we may have an apparent reflection point that is considerably beyond the end of the wires. Thus the distance from the end of the wires back to the first voltage minimum is much less than the quarter wave that otherwise might be expected.

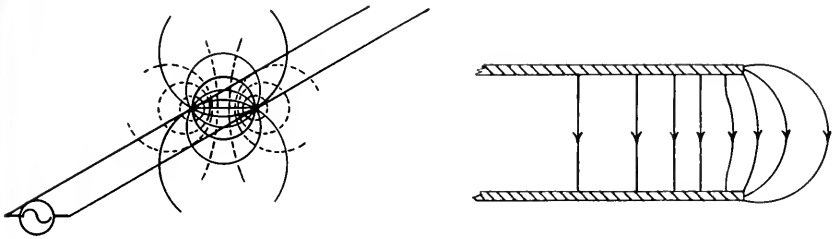

Fig. 6.2-7. (a) Transmission along a line open at the remote end. (b) Nature of reflection from open end.

It is generally true that processes of reflection in which fringing takes place are usually attended by considerable amounts of radiation. This suggests that in the process of reflection some of this extended wavepower detaches itself from the parent circuit and is lost. Experience shows that this lost power may be greatly enhanced by separating the two wires or by flaring their open ends. The so-called half-wave dipole, so familiar in ordinary radio, is but a transmission line in which the last quarter-wave length of each wire has been flared to an angle of 90 degrees. If we wish to minimize radiation, we follow a reverse procedure and reduce the spacing between the two parallel wires. This also reduces fringing, for we find that the measured distance from the ends of the wires to the first voltage minimum is now more nearly a quarter wave.

It is of interest to compare reflections taking place at the open end of a transmission line with those at a closed end. When a wave front becomes incident upon a perfect conductor, the electric force vanishes. At the same time, the lines of magnetic force, though effectively brought to rest, are

momentarily doubled in intensity. The energy is predominantly magnetic, and the type of reflection may be regarded as inductive. When the wave is reflected from the ideal open-end line, a reverse situation prevails. The lines of magnetic force momentarily vanish while lines of electric force, though brought to rest, are doubled in intensity. At this moment the energy is predominantly electrostatic, and the reflection may be considered as being capacitive.

When a line is terminated in a sheet of metal of good conductivity such as copper or silver, reflection is almost perfect. If the sheet is a poor conductor such as lead or German silver, most of the incident power will still be reflected; but if a semi-conductor, such as carbon, is used as a reflector, a perceptible amount of the incident power will be absorbed. It is interesting also that the penetration into all metals at the time of reflection is very slight, for relatively thin sheets seem to serve almost as well as thick plates. It is therefore possible to use as reflectors extremely simple and inexpensive materials, for example, foils or electrically deposited films fastened to a cheaper material such as wood.[7]

A more general study of reflections on transmission lines shows that the examples cited previously are special cases of a very general subject. Not only may there be reflections from the open and closed ends of a transmission line, but there may be reflections also when the line is terminated in an inductance, in a capacitance, or in a resistance. Details concerning the reflections that may be observed from various combinations of these three impedances are discussed in connection with Fig. 3.6-3. The outstanding results of these discussions may be summarized for the ideal case as follows:

1. A pure inductance (positive reactance) connected at the end of a transmission line always leads to a reflection coefficient having a magnitude of unity. The standing wave resulting from this reflection will be characterized by the following: (a) If the terminating inductance is infinitely large (reactance of positive infinity), the reflection will be identical with that from an ideal open-end line, and the distance to the nearest voltage minimum will be a quarter wave. [See Fig. 3.6-3(a).] (b) If the inductance is finite but very large, the distance to the nearest voltage minimum, as measured toward the generator, will be somewhat greater than a quarter wave. [See Fig. 3.6-3(b).] (c) If the inductance is reduced progressively toward zero (reactance zero), the distance to the same voltage minimum will approach one-half wavelength. In this limiting case, another voltage minimum will appear at the end of the line. [See Fig. 3.6-3(c) and 3.6-3(d).]

2. A pure capacitance (negative reactance) connected at the end of a

[7] One convenient and inexpensive form of reflector is a kind of building paper coated with copper or aluminum foil. Moderately good reflectors can also be made by covering wood with a special paint containing finely divided silver in suspension (Du Pont's 4817). Most aluminum paints are unsatisfactory for this purpose.

transmission line also leads to a reflection coefficient having a magnitude of unity. In this case, the resulting standing wave will be characterized as follows: (a) If the capacitance is zero, (reactance equal to minus infinity), the reflection will correspond to that from the open end of a transmission line, and a voltage minimum will be found at a distance of a quarter wave from the end. [See Fig. 3.6–3(g).] (b) If the capacitance is increased from zero to a small finite value, the distance to the nearest voltage minimum will be somewhat less than a quarter wave. [See Fig. 3.6-3(f).] (c) If the capacitance is increased progressively toward infinity (reactance zero), the distance to the nearest voltage minimum will approach zero. [See Figs. 3.6-3(e) and 3.6-3(d).] The limiting condition, in which the terminating capacitance is zero, is comparable with that in which the termination is an infinitely large inductance.

3. If a pure resistance is connected at the end of a transmission line, the magnitude of the reflection coefficient varies with the resistance chosen. The relations are such that: (a) If the terminating resistance is infinite, the magnitude of the reflection coefficient will be unity and its sign will be positive. [See Fig. 3.6-3(h).] (b) If the terminating resistance approaches the characteristic impedance of the line, the distance to the nearest voltage minimum will remain constant, but the magnitude of the reflection coefficient will approach zero. [See Figs. 3.6-3(i) and 3.6-3(j).] (c) If the terminating resistance is made less than characteristic impedance, the sign of the reflection coefficient will be reversed, and, as the terminating resistance approaches zero, its magnitude will approach unity. [See Figs. 3.6-3(k) and 3.6-3(l).]

When the terminating resistance is infinite, the reflection is comparable with that in an ideal open-end line, and the nearest voltage minimum will be found at a distance of a quarter wave. When the terminating resistance is zero, the reflection is comparable with that in a closed-end line, and the voltage minimum will appear at the end of the line and also at a point one-half wave closer to the generator. If the line is terminated in a pure resistance of intermediate value, the voltage minima of such standing waves as may be present will be found at the end of the line for all values of the resistance that are less than characteristic impedance and a quarter wave removed from the end of the line for all values greater than characteristic impedance. When the terminating resistance equals characteristic impedance, there is no standing wave.

If, instead of terminating the line considered above in an inductance coil or in a capacitance or a resistance, we assume that it continues indefinitely into a mass of material having either a conductivity or a dielectric constant different from that of air, similar reflections may take place at the surface. A particular example is shown in Fig. 6.2-8. In general, a part of the wave-power arriving at the surface will be reflected and a part will be transmitted.

One may picture a portion of the Faraday tubes of force turned back at the interface while the remainder continue into the second medium. If one were to reverse the direction of transmission and consider wavepower transmitted from the second medium back into the first, a similar partial reflection would be noted. In both cases the part turned back and returned to the source may be regarded as a reactive component since no energy is really lost. In a similar way, the transmitted component, since it is not returned to the source, may be regarded as a resistive or dissipative component.

If the medium into which wavepower is transmitted is a perfect insulator, the transmitted wave will continue indefinitely except as attenuated by the

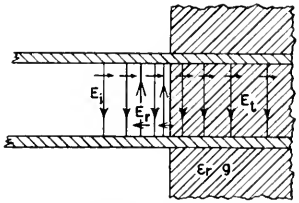

Fig. 6.2-8. Reflection and transmission of lines of force incidental to a change of medium along a transmission line.

wires along which it is guided. Its wavelength, λ, in the dielectric will be less than the wavelength, λ_0, in air as expressed by the relation

$$\lambda = \frac{\lambda_0}{\sqrt{\epsilon_r}}$$

If the second medium is somewhat conducting, the wave will be further attenuated, the rate of attenuation being related in a rather complicated way not only to the conductivity of the second medium but to its dielectric constant and permeability as well. Thus far in microwave practice, little practical use has been made of materials having permeabilities very different from unity. However, considerable use has been made of materials having various dielectric constants, ϵ_r, and conductivities, g. Sometimes these take the form of plates placed across a waveguide transmission line. Examples will appear in Section 9.8.

If a thin sheet of insulating material having a dielectric constant, ϵ_r, and conductivity of zero is placed across a two-wire transmission line, the percentage of power reflected is given approximately by

$$q_W = \frac{\pi t}{\lambda_0} (\epsilon_r - 1) \qquad (6.2\text{–}5)$$

A thin sheet of this kind is approximated when wires carrying very high frequencies pass through the glass walls of a vacuum tube. If the glass

thickness, t, is small compared with the wavelength in air, λ_0, the power reflected by the glass envelope will likewise be small.

Sometimes it is not feasible to reduce the wall thickness sufficiently to avoid serious reflections. In these instances it may be possible to make the thickness one-half wavelength as measured in glass whereupon the wave reflected from one face of the plate will be approximately equal in amplitude to that from the other face and, since they are separated by one-half wavelength, they tend to cancel.

Another case of practical interest is that in which the line is terminated in a plate of very special dielectric constant ϵ_r, conductivity g_1, and thickness t. This is followed by a second plate of nearly infinite conductivity. This arrangement is shown in longitudinal section in Fig. 6.2-9. By a proper choice of constants, the combination may be made a good absorber of wave-

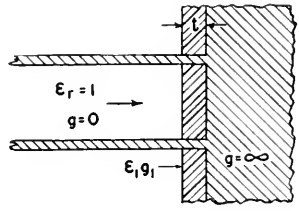

Fig. 6.2-9. A transmission line terminated in a conductor coated with a special materia such that all of the incident wave power is absorbed.

power. It will therefore be substantially reflectionless It may be shown that to satisfy this requirement

$$\lambda_0 = \frac{\sqrt{\epsilon_r}}{15\pi g_1} \tag{6.2-6}$$

and

$$t = \frac{1}{60\pi g_1(2n - 1)} = \frac{\lambda_0}{4\sqrt{\epsilon_r}\,(2n - 1)} \tag{6.2-7}$$

where n is any integer. One common example is that in which $n = 0$. The plate is then a quarter wave thick as measured in the medium.[8] A reflectionless plate of this kind when placed at the end of a transmission line appears to the source as though the line were terminated in its characteristic impedance. Devices incorporating this principle are sometimes used as match terminators for waveguides.[9]

[8] A more complete discussion of this problem was published in 1938 by G. W. O. Howe, "Reflection and Absorption of Electromagnetic Waves by Dielectric Strata," *Wireless Engr.*, Vol. 15, pp 593–595, November 1938.

[9] Plates of this kind may be made very simply by mixing carbon with plaster in varying proportions until the right combination is reached.

When a two-wire transmission line assumes the coaxial form, the lines of electric force are radial and lines of magnetic force are coaxial circles. The directions of these two components obey the right-hand rule. (See Fig. 6.2-10.) Since the wave configuration is completely enclosed except for a small exposure at each end, radiation from this type of line can be made very small.

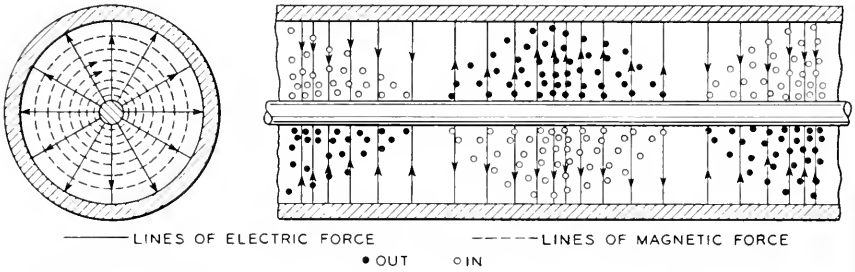

Fig. 6.2-10. Arrangement of lines of electric and magnetic force associated with transmission along a coaxial arrangement of conductors.

6.3 Radiation

Electromagnetic waves, including both light and radio waves, are not unlike the waves that are guided along wire lines. Their difference is largely a matter of environment. In one case they are attached to wires while in the other they have presumably detached themselves from some configuration of conductors and are spreading indefinitely into surrounding space. We shall present in this section one of several possible pictures of the launching of radio waves from a transmission line. Like other verbal pictures drawn in this chapter, it should be regarded as highly qualitative.

Assume a two-wire line with one end flared as shown in Fig. 6.3-1. If at some point to the left there is a source of wavepower, there will flow from left to right along the line a sinusoidal distribution of lines of electric and magnetic force not unlike that shown in Fig. 6.2-7. In order to simplify our illustration, we shall single out for examination two representative lines of electric force *a b* and *c d* located a half wave apart. It is understood, of course, that there are present many other lines both before and behind those represented. Also there are lines of magnetic force at right angles to the electric force. As time progresses each element of length of the line of force *a b* moves laterally with the velocity of light. In the region where the wires are parallel, it remains straight but, upon reaching the flared section, its two ends fall behind the central section, thereby forming a curve as shown in Fig. 6.3-1(c). As this line of force moves to the end of the flared section [Fig. 6.3-1(d)], its successor *c d* follows one-half wavelength behind.

Because of the property of inertia with which all lines of force are assumed to be endowed, the central section of *a–b*, which is already greatly extended due to curvature, continues in motion for some time after the two ends, attached to the conductors, have come to rest. The result is shown approximately by Fig. 6.3-1(e). An instant later and perhaps after the two ends of line of force *a–b* have started on their return journey, the line of force *c–d* approaches sufficiently close to *a–b* that a coalescence ensues [Fig. 6.3-1(f)]. An instant later fission takes place as illustrated in Fig. 6.3-1(g), leaving a portion of the energy of each *a–b* and *c–d* now shared by a radiated com-

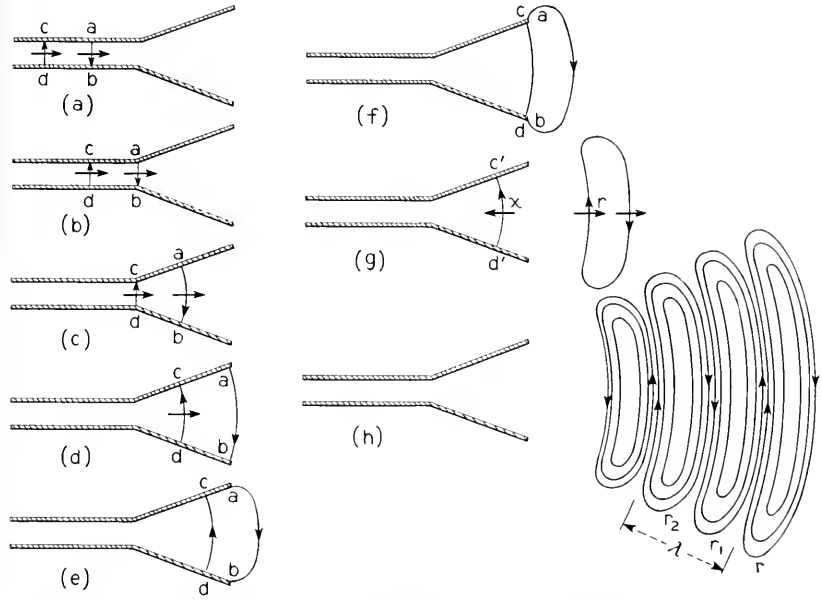

Fig. 6.3-1. Successive epochs in a highly idealized representation of radiation from the flared end of a transmission line.

ponent, *r*, and a reflected component, *x*. That the two components *r* and *x* should travel in opposite directions seems reasonable when it is noted that lines of electric force in *x* are in the same direction as in the adjacent portion of *r*. They may therefore be expected to repel. The first of these components, *r*, appears to the transmitter as though it were a resistance since it represents lost energy. The second, *x*, appears as a reactance since it represents energy returned to the transmitter. The radiated component, *r*, will be followed by other components r_1, r_2, etc., as represented in Fig. 6.3-1(h).

In the radiated wave front, the two components *E* and *H* are everywhere mutually perpendicular and in the same phase. Because the wave front

is curved, as shown in cross section in Fig. 6.3-2, the component Poynting vectors which specify the directions in which energy is flowing will be slightly divergent. As a result, only a portion of the total wavepower will proceed in the preferred direction. It follows that, for best directivity, the emitted wave front should be substantially plane, and the lines of force should be as nearly straight as possible. There is shown in Fig. 6.3-3 a series of configura-

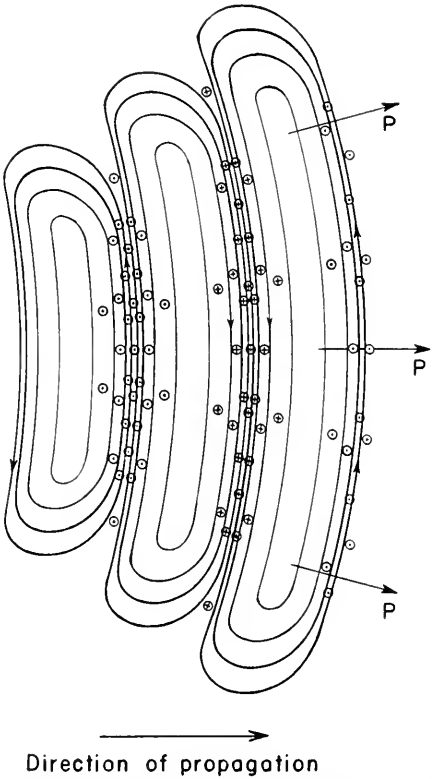

Fig. 6.3-2. Cross section of electromagnetic waves radiated from the flared end of a transmission line. Lines of electric force lie in the plane of the illustration; lines of magnetic force are perpendicular to the illustration while the flow of power is along the divergent arrows P.

tions based partly on speculation and partly on deductions from Huygens' principle. They illustrate in a rough way how, by increasing the aperture between the two wires of the elementary radiator, we may make the individual component Poynting vectors more nearly parallel.[10]

[10] Figure 6.3-3 has been greatly oversimplified. Experiment shows that, to achieve the result desired, the angle between the two wires of Fig. 6.3-3 must be smaller for larger apertures than for small apertures.

Thus far, we have restricted our considerations to directivity in the plane of the two conductors (vertical plane as here assumed). Experiment shows that, in the plane perpendicular to that illustrated, the directivity from a single pair of wires is slight. However, we may obtain additional directivity by increasing the horizontal aperture. One method of accomplishing this result is to array, at rather closely spaced intervals, identical elementary radiators each of the kind just described. [See Fig. 6.3-4(a).] An infinite number of these elements infinitesimally spaced become two parallel plates as shown in Fig. 6.3-4(b). If metal plates are now attached at the right and left

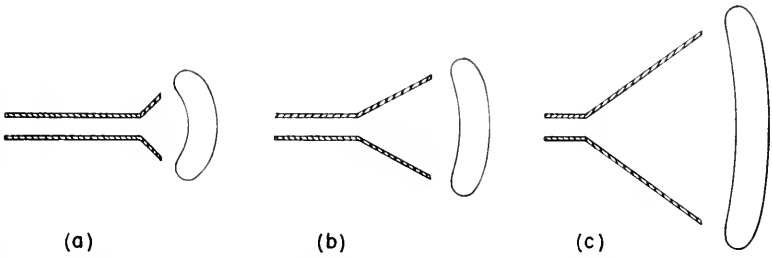

Fig. 6.3-3. Illustrating how radiating systems of large aperture may give rise to wave fronts of large radius of curvature and hence lead to increased directivity.

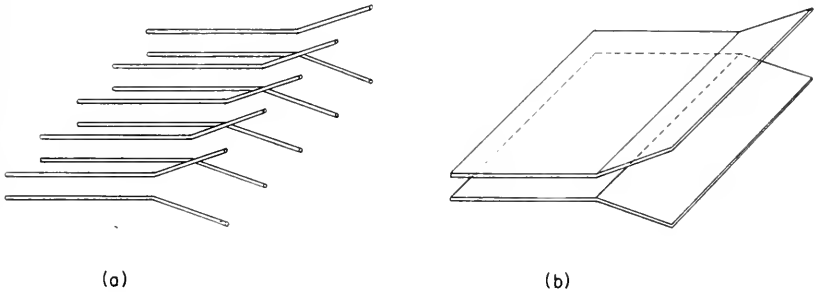

Fig. 6.3-4. Alternate ways by which the aperture of a flared transmission line radiator may be increased.

sides, the resulting configuration will become a waveguide horn. As a general rule, the larger the area of aperture, the more directive will be the antenna. The highly schematic array shown in Fig. 6.3-4(a) is introduced for illustrative purposes only. It is not one of the preferred forms used in microwave work. More practicable forms will be found in Chapter X.

The wave model shown in Fig. 6.3-2 conveys but a portion of the known facts about a radiated wave. A more accurate model is shown in skeleton form in Fig. 6.3-5. It is assumed that the transmitted wave has been launched with about equal directivity in the two principal planes and that the ob-

server is looking into one-half of a cut-away section of the total configuration. In the complete configuration, the individual lines of electric force (solid lines) and magnetic force (dotted lines) form closed loops, thereby producing in each half-wave interval a packet of energy. The stream of projected energy from an antenna is, according to this view, a series of these packets one behind the other moving along the major axis of transmission. At the transmitter each packet may have lateral dimensions that are only slightly greater than the corresponding dimensions of the radiating antenna; but, since the packet has curvature and since propagation is radial, the packet spreads as it progresses so that at the distant receiver it may be very large indeed.

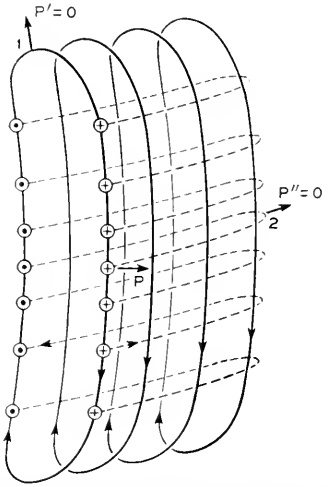

Fig. 6.3-5. Highly idealized representation of a wave-packet radiated by a typical microwave source. One half of the total packet is assumed to be cut away.

Around the edge of each packet there is a region where the relationship between the vectors E, H, and v is rather involved. For example, in the vicinity of point 1 in Fig. 6.3-5, there is a substantial component of E but at this point the vector H is zero and accordingly the Poynting vector P' at that point is also zero. (See Equation 6.1–4.) In a similar way there may be in the vicinity of point 2 a substantial component of magnetic force **H**; but, since at this point the electric force is substantially zero, we conclude that the Poynting vector P'' is again zero and again no power is propagated.[11]

[11] The peculiar edge effects noted may be regarded as a result of a kind of wave interference not unlike that prevailing in the regions of minimum E and H in the case of standing waves as discussed in Section 6.3. A similar kind of wave interference is cited in Section 6.5 to account for regions of low E and H in transmission along a waveguide.

The sharpest radio beams now in general use are only a few tenths of a degree across. We conclude that for these sharp beams a small but nevertheless appreciable curvature remains in the radiated wave packet. This means that, when the wave front has arrived at a distant receiver, it is still many times larger than any receiving antenna it may be practicable to construct, and accordingly the latter can intercept but a small portion of the total advancing wavepower. This implies a considerable loss of power, which is indeed the case.

In the process of radio reception, one may think of the antenna structure as a device that cuts from the advancing wave front a segment of wavepower which it subsequently guides, preferably without reflection, to the first stages of a nearby receiver. To be efficient, the wavepower intercepted should be large. This, in turn, calls for a receiving antenna of considerable area. It will be remembered that a large aperture was also a necessary feature for high directivity at the transmitter. This is consistent with the accepted view that the processes of reception and transmission through an antenna are entirely correlative and that a good transmitting antenna is a good receiving antenna and vice versa. The directive properties of an antenna are sometimes specified in terms of its *effective area.* (See Section 10.0.)

The term *uniform plane wave* is a highly idealized entity assumed in many problems for purposes of simplicity but never quite attained in practice. In an idealized wave front, the electric and magnetic components E and H are not only everywhere mutually perpendicular but both components are exclusively transverse. That is, there is no component of either E or H in the direction of propagation. Such a wave belongs to a class known as *transverse electromagnetic waves* (TEM). These may be compared with others, to be described later, known as *transverse electric waves* (TE) and *transverse magnetic* (TM) *waves.* Waves guided along parallel conductors are also TEM waves, but except in the case of infinitely large conductors they are not *uniform plane waves.*

6.4 Reflection of Space Waves from a Metal Surface

One of the early triumphs of the electromagnetic theory was its ability to account satisfactorily for the reflection and refraction of light. This theory was so general as to include not only a wide range of wavelengths but also a wide range of surfaces as well. According to this theory, reflections may occur whenever electromagnetic waves encounter a discontinuity. This may happen, for example, when waves fall on a sheet of metal, in which case the discontinuity is due to the sudden change in conductivity. Reflection may also occur when waves are incident on a thick slab of glass or hard rubber, in which case reflection is due to a sud-

den change in dielectric constant.[12] Similar reflections may theoretically take place also at an interface where the permeability of the medium changes suddenly. The case in which there is a change of conductivity has an important bearing on waveguide transmission. It will therefore be discussed in considerable detail.

Assume a plane wave incident obliquely upon a conducting surface as shown in Fig. 6.4-1. The line along which the wave is progressing (wave-normal) is referred to as the *incident ray*. It intersects the conducting surface or interface at a point O and makes an angle θ with the perpendicular OZ. After reflection, the normal to the new wave wave front makes an angle θ' with the perpendicular OZ. This second wave-normal is known as the

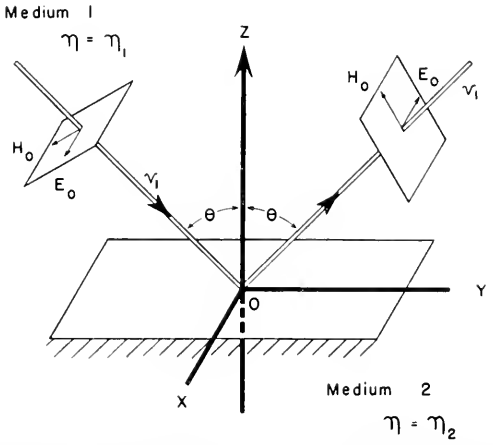

Fig. 6.4-1. Reflection at oblique incidence from a metal plate for the particular case where the electric vector is perpendicular to the plane of incidence.

reflected ray, and its angle with the perpendicular OZ is known as the *angle of reflection*. The plane containing the incident ray and the perpendicular OZ is known as the *plane of incidence*. The incident and reflected rays lie in the same plane, and their corresponding angles of incidence and reflection are numerically equal.

In problems of oblique incidence there are two cases of interest, depending on whether the electric or the magnetic component lies in the plane of incidence. For our particular purpose, the second of these two cases is of special interest and it will therefore be discussed in considerable detail. The vector relations corresponding to this case are shown in Fig. 6.4-1.

[12] For a more general discussion of the electromagnetic theory of reflection: L. Page and N. I. Adams, "Principles of Electricity," D. Van Nostrand Co., Inc., pp 569–575, New York 1931. R. I. Sarbacher and W. A. Edson, "Hyper and Ultra-high Frequency Engineering," John Wiley & Sons, Inc., pp 105–116, New York 1943.

Included are the relative directions of E and H both before and after reflection.

In Fig. 6.4-2 there are shown in cross section representative lines of electric force in an advancing plane wave front. They are numbered respectively 1, 2, 3, 4, 5, 6, and 7. Each individual figure [(a), (b), (c), etc.] represents a succeeding period of time. We shall assume that the particular wave front singled out for illustration represents the crest of a wave. Both ahead and behind this crest there are located alternately at half-wave intervals other crests and hollows, and their respective lines of force alternate in direction. Each line of force in the wave front is assumed to be moving in a direction indicated by the vector v. It is furthermore assumed that there is also present a magnetic component, indicated by the dotted vector H that is perpendicular to E and also to v. The vectors v and H must of course be so directed as to be in keeping with the right-hand or cork-screw rule, both before reflection and after reflection. Also at the point of incidence the tangential electric force must be zero. To account for this, we assume that as each line of electric force moves up to the conducting plane it is reversed in direction, thereby making on the average as many lines of electric force at the surface directed toward the observer as directed away from the observer. Consider, for example, lines of force 3 and 5, 2 and 6, and 1 and 7, in Fig. 6.4-2(c).

Associated with these two components of electric force which, let us say, are E and E', there are two components of magnetic force H and H'. These may be specified by $\mathbf{H} = \epsilon(\mathbf{v} \times \mathbf{E})$, each of which at the interface may be resolved into two components shown in Fig. 6.4-3 as $H = H_\perp + H_\parallel$ at the left and $H_\perp' = -H_\parallel'$ at the right. Combining these four vectors, assuming reflection to be perfect, we find that at the interface $H_\perp - H_\perp' = 0$ and $H_\parallel - (-H_\parallel') = 2H$, giving as an over-all result: (1) the electric force at the interface is everywhere zero; (2) the vertical component of the magnetic force at this point is also zero; and (3) the tangential component of the magnetic force at the interface is $2H$.

The peculiar configuration that resides close to the metal boundary is propagated to the right as a kind of magnetic wave. It has rather interesting properties which will become more evident by referring again to Fig. 6.4-2. Two conclusions may be drawn from this figure, depending on the point of view assumed. To a myopic observer located at the interface and unable to see far beyond the point p and unable to distinguish one line of force from another, the advancing wave front would look like a configuration of amplitude $H_\parallel = 2H$ and $E_\parallel = 0$ moving parallel to the interface with velocity $v_z = v/\sin\theta$. To this observer the apparent velocity would increase as θ becomes progressively smaller until, at perpendicular incidence, v_z would approach infinity. These results follow from the geo-

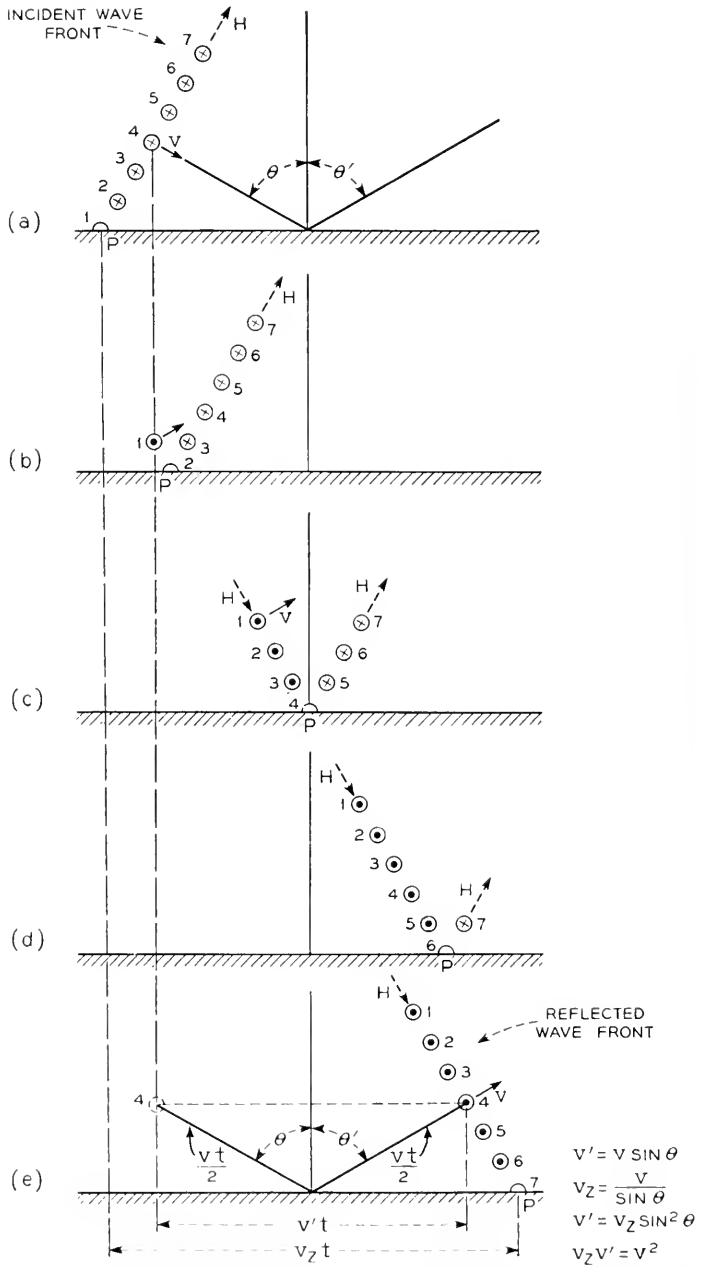

Fig. 6.4-2. Successive steps in the reflection of a single plane wave front by a metal plate.

metrical relations shown in the lower part of Fig. 6.4-2. Phenomena similar to this are sometimes observed when water waves, coming in from the ocean, break upon the beach. If the approach is nearly perpendicular, the point at which the wave breaks may proceed along the beach at a phenomenal speed. A similar effect may be produced by holding at arm's length a pair of scissors and observing the point of intersection as the blades are showly closed. A relatively slow motion of the blades leads to a rather rapid motion of the point of intersection.

Since, in the case of incident waves, the apparent velocity is $v_z = v/\sin\theta$, the corresponding wavelength is $\lambda_z = \lambda/\sin\theta$. Both quantities play an important part in the picture of waveguide transmission to be drawn later. In particular, the apparent velocity v_z will prove to be identical with a quantity known as *phase velocity*.

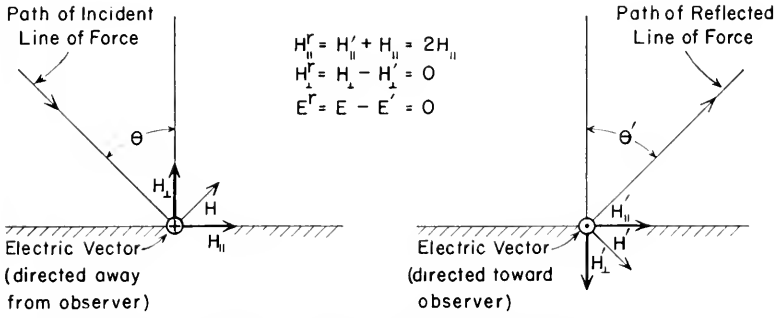

Fig. 6.4-3. Relationship between various components of E and H before and after reflection by a metal plate.

A second observer located at the interface, shown in Fig. 6.4-2, endowed with better vision and able to single out particular lines of force may obtain a somewhat different view of reflection. If he observes a particular line of force such as (4) in Fig. 6.4-2 for the considerable period of time, t, required for it to approach the conducting interface [Figs. (a) to (c)] and recede to a comparable distance [Figs. (c) to (e)], he will note that, whereas the line of force has really traveled a total distance vt, its effective progress parallel to the interface has been $v't = vt \sin\theta$. (See geometrical relations in lower part of Fig. 6.4-2.) This provides another kind of velocity ($v' = v \sin\theta$) known as *group velocity*. It is the effective velocity with which energy is propagated parallel to the metal surface. It approaches zero at perpendicular incidence. It will be observed that

$$v' = v_z \sin^2\theta$$

and

$$v'v_z = v^2 \tag{6.4-1}$$

Group velocity also plays an important part in waveguide transmission.

6.5 Waveguide Transmission

It was pointed out in an earlier chapter that each of the various configurations observed in waveguides may be considered as the resultant of a series of plane waves each traveling with a velocity characteristic of the medium inside, all multiply reflected between opposite walls. In the case of certain of these waves, this equivalence may not be readily obvious, but for the dominant mode in a rectangular guide, which is one of the more important practical cases, it is relatively simple. It also happens that the analysis of such waves throws considerable light on the nature of guided waves, and furthermore it enables us to deduce many of the useful relations used in waveguide practice—relations that might otherwise call for rather complicated mathematical analysis.

It is assumed in Fig. 6.5-1 that we are viewing, in longitudinal section and at successive intervals of time, a hollow rectangular pipe having transverse dimensions of a and b measured along the x and y axes respectively. In this case the illustration is in the xz plane. It is further assumed that the electric force lies perpendicular to the larger dimension a and is consequently perpendicular to the plane of the illustrations. We assume in Fig. 6.5-1 (a) a particular plane wave front 1, perhaps a crest, that has recently entered the guide from below. Let us say that its velocity is $v = v_a/\sqrt{\mu_r \epsilon_r}$ and that is it so directed as to make an angle θ with the left-hand wall as shown.[13] Reflection at the left-hand wall will therefore be identical with that already shown in Fig. 6.4-2. A portion of the wave front that has just previously undergone reflection is shown immediately below at 2 in Fig. 6.5-1(a). We assume further that this front is made up of lines of electric force perpendicular to the illustration together with associated lines of magnetic force lying in the plane of the illustration. It will be obvious presently that, like the case of reflection from a single conducting sheet discussed in the previous section, we may obtain two rather different pictures of what takes place within the guide, depending on whether we fix our attention on the configuration as a whole or on some particular line of force which we may identify and follow through a considerable interval of time. We shall first consider the configuration as a whole.

[13] It is to be noted that the angle ϕ which the wave front makes with the metal wall is equal to the angle which the wave-normal (ray) makes with the perpendicular to the metal wall.

We show in Fig. 6.5-1(b) the same wave front shown in Fig. 6.5-1(a) but at an epoch later—after it has progressed a considerable distance along the guide. We now find the reflected portion 2 complete and a new portion

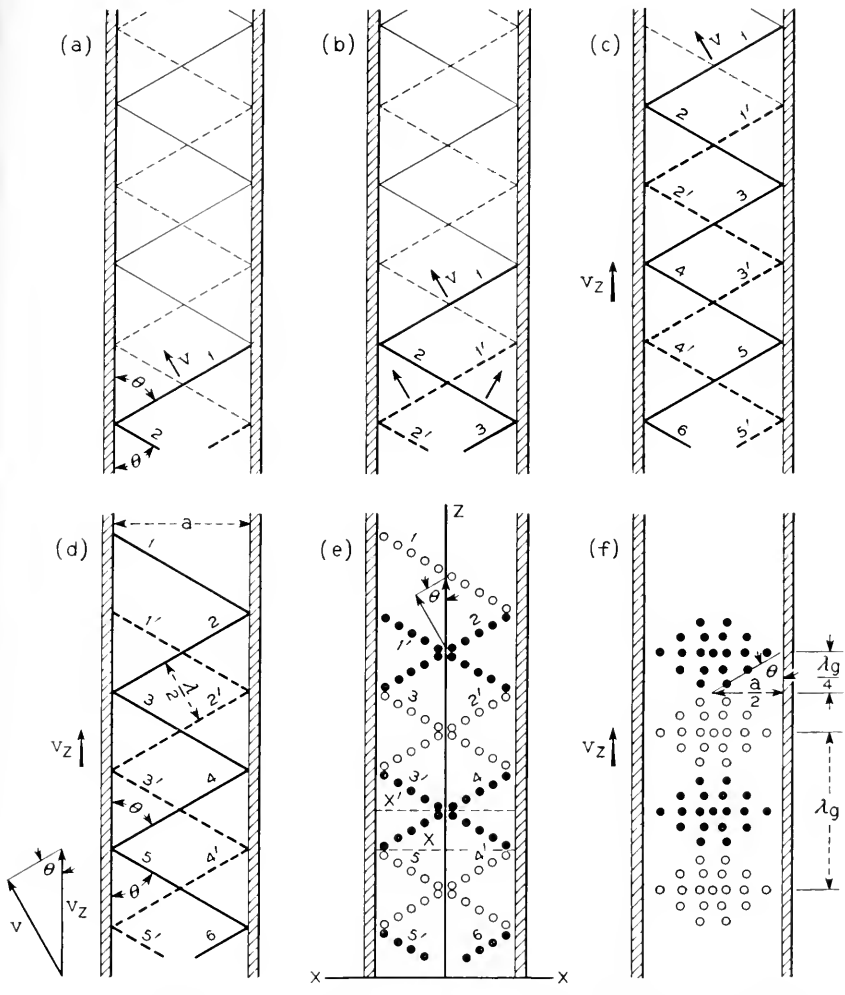

Fig. 6.5-1. The propagation of a multiply reflected wave front between two metal plates [Figs. (a)–(d)] is equivalent to the transmission of a TE wave parallel to the two plates. [Fig. (f)].

3 about to enter the guide. Following wave front 1 and at a distance of one-half wave behind, we find, shown dotted, the "hollow" of the wave. This we shall designate by the numeral 1′. We find here also a new portion of the "hollow" 2′ that has just undergone reflection.

In Fig. 6.5-1(c) and again in Fig. 6.5-1(d) we find successive positions of these same wave fronts as they have moved forward in the guide. We may, if we like, think of these fronts as discrete waves moving zig-zag through the guide or as a single large wave front folded repeatedly back upon itself. Fixing our attention for the moment on Fig. 6.5-1(d), we observe that the velocity v at which any point of incidence of the wave front (say at point 5) moves along the guide is given by the relation

$$v_z = \frac{v}{\sin \theta}$$

This particular velocity v_z is the phase velocity of the wave as seen by a myopic observer located near a lateral wall of the guide.

Referring again to Fig. 6.5-1(d) and fixing our attention on the geometrical relation between the wavelength λ and the width of the guide a, we may construct a right triangle with $\lambda/2$ and a as sides and show that

$$\cos \theta = \frac{\lambda}{2a} \tag{6.5-1}$$

and since

$$\sin \theta = \sqrt{1 - \cos^2 \theta} \tag{6.5-2}$$

$$\sin \theta = \sqrt{1 - \left(\frac{\lambda}{2a}\right)^2} \tag{6.5-3}$$

and

$$v_z = \frac{v}{\sqrt{1 - \left(\frac{\lambda}{2a}\right)^2}} \tag{6.5-4}$$

This says that for very large guides, that is, $\lambda < 2a$, $v_z \doteq v$, but as λ approaches $2a$, v_z approaches infinity. The particular case where $\lambda = 2a$ and $v_z = \infty$ is referred to as the *cut-off condition*. At cut-off, it would appear that the individual waves approach the wall at perpendicular incidence and a kind of resonance between opposite walls prevails. At wavelengths greater than cut-off no appreciable amount of power is propagated through the guide.

The particular value of wavelength measured in air, corresponding to cut-off, is referred to as the *critical* or *cut-off wavelength* and is designated thus: $\lambda_c = 2a$. The corresponding frequency is similarly known as the *critical* or *cut-off frequency* and it is designated thus: $f_c = v/\lambda_c$. It is sometimes convenient to designate the ratio of the operating wavelength to the critical wavelength by the symbol ν. From Equation 6.5-4 it follows that

$$\frac{v_z}{v} = \frac{1}{\sqrt{1 - \left(\frac{\lambda}{\lambda_c}\right)^2}} = \frac{1}{\sqrt{1 - \left(\frac{f_c}{f}\right)^2}} = \frac{1}{\sqrt{1 - \nu^2}} \qquad (6.5\text{–}5)$$

Referring to Fig. 6.5-1(a) we have indicated that the wave front 1 is made up of lines of electric force directed through the plane of the illustration and hence away from the observer. There are, of course, lines of magnetic force and also other lines of electric force both ahead and behind the wave front drawn, but these have purposely been omitted in order to simplify the illustration. If we were to take the magnetic force into consideration we would find as in Fig. 6.4-2 that, at the reflecting surface, a tangential component only is present and its magnitude is twice that of the magnetic component of the incident wave.

In the discussion of reflection of plane waves in the previous section, it was also pointed out that the act of reflecting a wave reverses the direction of the electric force. Applying this principle to the case at hand, we see that if the electric force is directed downward in the section of wavefront 1 of Fig. 6.5-1(a), it will be directed upward in 2. Carrying this idea forward to Fig. 6.5-1(e) we find that in fronts 1, 2, 3, etc., which we rather arbitrarily called crests, the electric vector alternates in direction as shown by the open and solid circles. Likewise the direction of the electric vector alternates in the fronts designated as 1′, 2′, and 3′, but in this case they are respectively opposite in direction to 1, 2, and 3. Continuing to fix our attention on Fig. 6.5-1(e), it will be observed that the direction of lines of force is the same in 1′ and 2, in 2′ and 3, and in 3′ and 4, indefinitely along the entire length of the guide. Thus there are regularly spaced regions along the length of the guide where the electric vector is directed toward the observer alternating with other regions where the electric vector is directed away from the observer. Between the two are still other regions where the respective component vectors are oppositely directed and hence their sum may be zero.

Adding the foregoing effects, bearing in mind that there are lines of force both ahead and behind the highly simplified wave fronts shown, we have a new wave configuration moving parallel to the main axis of the guide with a phase velocity v_z as suggested by Fig. 6.5-1(f). Examining more carefully the wave interference that is here taking place, it becomes evident that if we pass laterally across the guide along the line x in Fig. 6.5-1(e) the instantaneous value of the resultant electric vector as shown is everywhere zero. On the other hand, if we cross the guide along a parallel line x', the electric vector varies sinusoidally beginning at zero at either wall and reaching a maximum in the middle of the guide. It will be observed that if we pass along the major axis z of the guide the electric vector at

any instant again varies sinusoidally with distance. However, at the boundary of the guide the resultant electric vector is everywhere zero. Since there was no component of the electric force lying along the axis z of the guide in the component waves that gave rise to this configuration, there can be no such component in the resultant. Waves in which the electric vector is exclusively transverse are known as *transverse electric*, or TE, *waves*.

A complete account of transmission of this kind should include, of course, a consideration of the lines of magnetic force. From Fig. 6.4-3 it is evident that, at the point of reflection of the component plane wave on the guide wall, there are two components of magnetic force $H_\perp$ and $H_\parallel$ in both the incident and reflected waves. When these are added, the resultant of the transverse magnetic force, like that of the electric force, differs at different points in the guides. Following alone the line x', it is found that for the particular condition here assumed, the magnetic force is zero at each wall increasing sinusoidally to a maximum midway between. At this point the magnetic component is entirely transverse. Following along the line x, it will be found that the magnetic vector is a maximum near each wall decreasing cosinusoidally to zero in the middle. It is of particular interest that, at the wall of the guide, the magnetic component lies parallel to the axis. Magnetic lines of force are, in this type of wave, closed loops, whereas lines of electric force merely extend from the upper to the lower walls of the guide. The arrangement of lines of electric and magnetic force in this type of wave is shown in Fig. 5.2-1. The quantitative relationships between the various components of E and H are specified more definitely by Equation 5.2–1. The significance of the wavelength λ_g of this new configuration will be obvious from Fig. 6.5-1(f).

There are certain useful results that follow from Fig. 6.5-1(f). It may be seen from the triangle there shown that

$$\frac{\lambda_g}{4} = \frac{a}{2} \cot \theta \tag{6.5–6}$$

From Equations 6.5-1 and 6.5–3, it will also be seen that

$$\cot \theta = \frac{\cos \theta}{\sin \theta} = \frac{\lambda}{2a \sqrt{1 - \left(\frac{\lambda}{2a}\right)^2}} \tag{6.5–7}$$

Therefore

$$\lambda_g = \frac{\lambda}{\sqrt{1 - \left(\frac{\lambda}{2a}\right)^2}} = \frac{\lambda}{\sqrt{1 - \nu^2}} \tag{6.5–8}$$

Since $1/\sqrt{1 - \nu^2}$ is the ratio of the apparent wavelength in the guide to that in free space and since for hollow pipes it is greater than unity, it is sometimes referred to as the *stretching factor*. It appears frequently in quantitative expressions relating to waveguides. Since velocity is equal to the number of waves passing per second times the length of each wave, we have

$$v_z = \frac{v}{\sqrt{1 - \nu^2}} \tag{6.5-9}$$

This is equivalent to the relation shown as Equation 6.5–5.

A matter of special interest is the rate at which energy is propagated along the guide. For present purposes, it is convenient to regard a moving line of force and its associated magnetic force as a unit of propagated energy. A knowledge of the path followed by such a line of force will therefore shed light on the rate at which energy is propagated along a waveguide.

It was pointed out in connection with Equation 6.4–2 that, when a wave is incident obliquely upon a metal surface, the apparent phase of the wave progresses at a velocity v_z greater than the velocity of light v, but that the energy actually progresses parallel to the interface at a velocity v' less than the velocity of light. It was pointed out, too, that $v' = v \sin \theta = v_z \sin^2 \theta$. Because of multiple reflections between opposite walls of a waveguide, its *phase velocity* is identical with v_z. Also, because of these multiple reflections, energy being carried by these component plane waves follows a rather devious zig-zag path and will therefore progress along the axis of the guide at a relatively slow rate. This velocity which is known as the *group velocity* is idential with v' above. From relations already given, it will be seen that

$$v' = v\sqrt{1 - \nu^2} \tag{6.5-10}$$

also

$$v' = v_z(1 - \nu^2) \tag{6.5-11}$$

It will be apparent from this relation that, at cut-off, where $\nu = 1$, energy is propagated along the guide with zero velocity. This is consistent with the idea already set forth that, at cut-off, energy oscillates back and forth between opposite faces of the guide. As we leave cut-off and progress toward higher frequencies (shorter waves), the group velocity v' increases as the phase velocity v_z decreases, until, at extremely high frequencies, both approach the velocity v characteristic of the medium. This relationship is made more evident by Fig. 6.5-2.

Reviewing again the simple analysis just made, we find that the wave configuration that actually progresses along a conventional rectangular waveguide may be regarded as the result of interference of ordinary uni-

form plane waves multiply reflected between opposite walls of the guide. This viewpoint accounts for not only the distribution of the lines of force in the wave front but also for the velocity at which the phase progresses and the velocity at which energy is propagated. As we shall soon see, it accounts also for the rate of attenuation.

In the particular configuration just described the electric component is everywhere transverse, whereas the magnetic component may be either longitudinal or transverse, depending on the point in a guide at which observations are made. These waves are plane waves, but, since the elec-

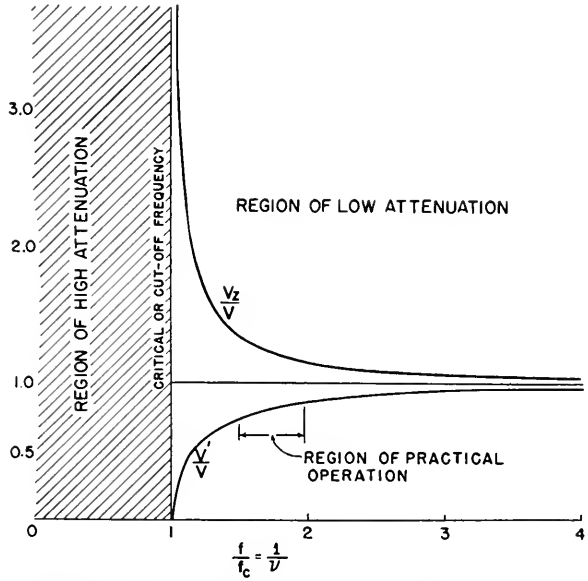

Fig. 6.5-2. Relative phase velocity v_z and group velocity v' for various conditions of operation of a waveguide.

tric intensity is not uniformly distributed over the wave front, they are not uniform plane waves.

The concept of multiply reflected waves provides a basis for calculating the attenuation in rectangular guides as was shown by John Kemp several years ago.[14] The procedure is outlined briefly below. The reader is referred to the published article for details.

There is shown in Fig. 6.5-3 a short section of hollow waveguide in which we imagine multiply reflected plane waves are propagated. We fix our attention on a zig-zag section cut from the guide and so directed that it

[14] John Kemp, "Electromagnetic Waves in Metal Tubes of Rectangular Cross-section," *Jour. I.E.E.*, Part III, Vol. 88, No. 3, pp 213–218, September 1941.

lies parallel to the direction of propagation of the elemental wave fronts. The top and bottom conductors so formed may be regarded as a uniform flat-conductor transmission line with oblique reflecting plates (sections of the side walls) spaced at regular intervals. Other transmission lines adjacent to that under consideration behave in exactly the same way as that singled out for examination and at the same time act as guard plates to insure that the lines of force so propagated remain straight.

It is clear that the attenuation in each elemental transmission line will be that incidental to losses in the upper and lower conductors plus the losses incidental to reflection at oblique incidence from the several reflecting

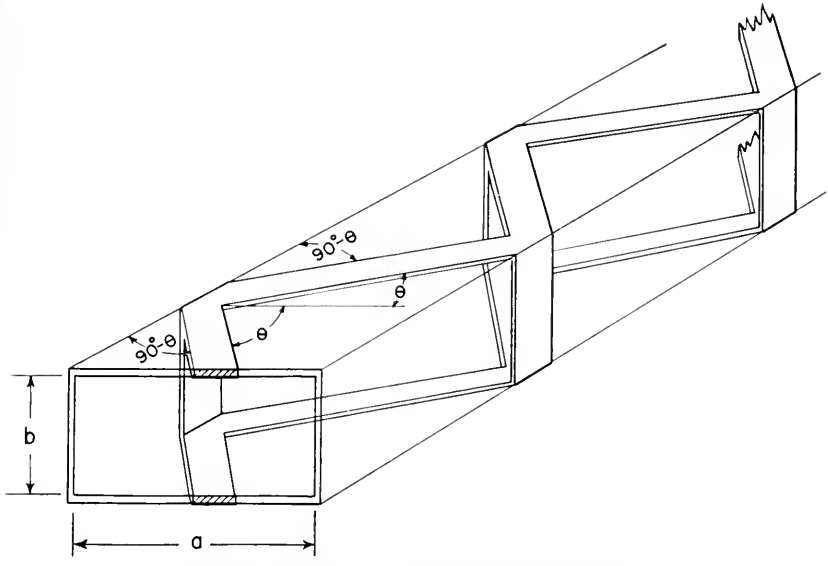

Fig. 6.5-3. Elementary transmission lines terminated periodically by reflecting plates which go to make up a rectangular waveguide.

plates. The total attenuation of the rectangular guide may then be found by summing up over a unit length of waveguide all of the elemental lines. This has been done with results that are equivalent to the corresponding equations given in Chapter V. The results are plotted in Fig. 6.5-4.

Certain characteristics of these curves may be readily accounted for. For instance, at cut-off ($\theta = 0$), both the number of unit reflection plates and the number of flat-plate transmission lines in a given length of waveguide will be infinite. As a result, the component attenuations arising in each of these two sources will likewise be infinite. As the frequency is increased above cut-off the angle θ will increase accordingly, leading thereby

to fewer side-wall reflections and to a shorter over-all length of zig-zag transmission line. Thus, in this frequency range, the attenuations contributed both by the side walls and by the top and bottom plates decrease with increasing frequency. Proceeding to frequencies far above cut-off, where θ approaches 90 degrees, there will not only be very few reflections but the over-all length of zig-zag line will approach as its limit a single, straight two-conductor line made up of the top and bottom plates alone. Thus the attenuation due to the side walls will approach zero and that due to the top and bottom plates will increase as the square root of the

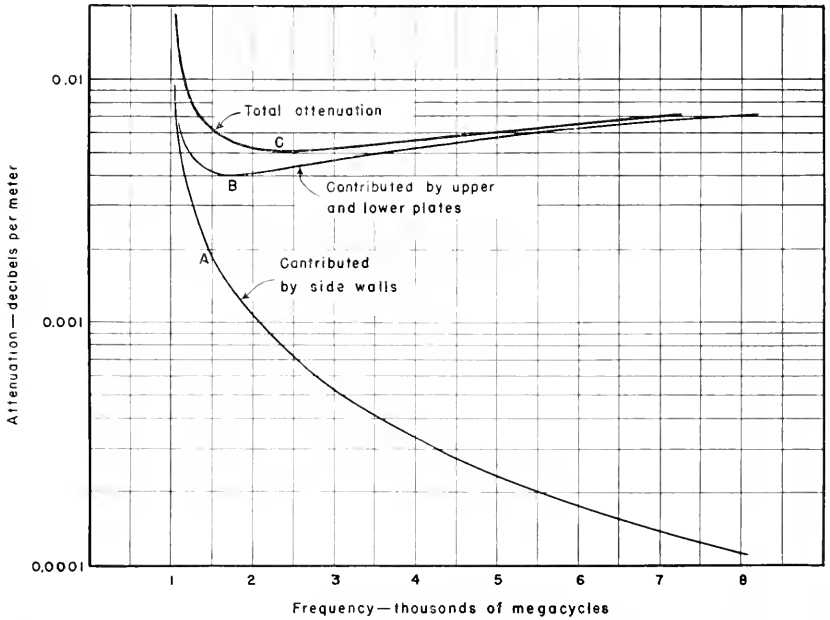

Fig. 6.5-4. Component attentuations contributed by the top and bottom plates and also the two side walls of a rectangular waveguide.

frequency. Since the attenuation contributed by the top and bottom plates first decreases but later increases with frequency, we may expect, between these two ranges, a region of minimum attenuation. The attenuations contributed by the upper and lower plates and also by the side walls of a 7.5 cm $\times$ 15 cm copper guide carrying the dominant mode have been calculated. The results have been plotted as curves A and B in Fig. 6.5-4. They follow the courses predicted by the preceding qualitative reasoning.

The fact that the reflection type of attenuation, such as is evident in the side walls above, decreases with frequency, suggests that, if a kind of wave-

guide could be devised where this type of attenuation alone exists, we could then operate the guide at extremely high frequencies and thereby obtain relatively low attenuations. This can, in effect, be done. It calls for a guide of circular cross section and a special configuration, known as

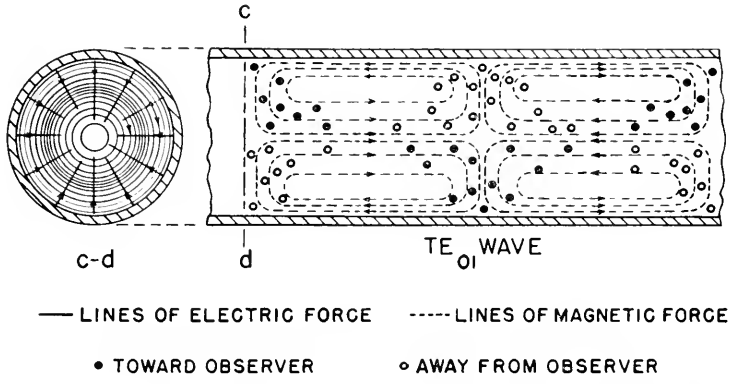

Fig. 6.5-5. The circular electric or TE_{01} configuration in a circular waveguide.

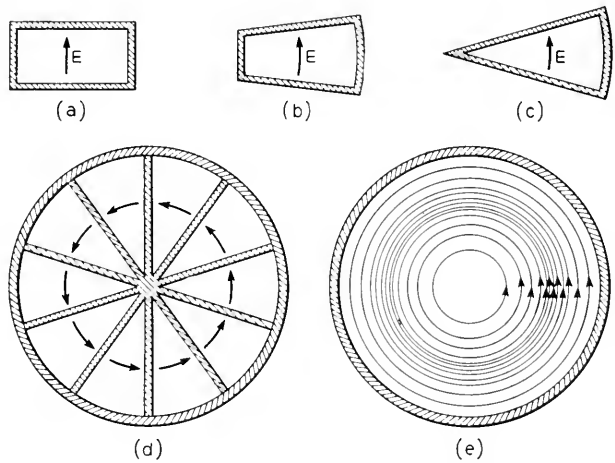

Fig. 6.5-6. Evolution of the circular-electric wave in a circular pipe from a dominant wave in a rectangular pipe.

the *circular-electric wave*. In this configuration, the resultant electric force is everywhere parallel to the conducting boundary as shown in Fig. 6.5-5.

That such a wave will lead to the interesting frequency characteristic noted is made more plausible by referring to Fig. 6.5-6 and its associated discussion. Figure 6.5-6(a) shows a conventional form of rectangular guide in which plane waves are multiply reflected from the two short sides.

In Fig. 6.5-6(b) the proportions of the guide have been altered somewhat, but since the lines of electric force are still perpendicular to the top and bottom plates, the guide may be expected to function substantially as before. At the most, some attenuation that previously originated in the left-hand side wall may now be transferred to the top and bottom walls. As a second step, we may extend the width of the top and bottom walls as shown in Fig. 6.5-6(c) until they intersect, thereby forming an arc-shaped guide. The attenuation now prevailing is evidently confined to the top and bottom walls and the right-hand wall. It is reasonable to assume that the side wall attenuation still decreases with frequency since incident lines of force are everywhere parallel to this wall. As a third step, we assemble as in Fig. 6.5-6(d) a number of identical arc-shaped guides to form a composite circular guide with radial partitions. If, finally, we imagine the radial partitions removed as in Fig. 6.5-6(e), the resulting configuration will not be altered and we shall have removed the component of attenuation attributable to the top and bottom walls leaving only the component of attenuation attributable to the one side wall, which, as we have pointed out, becomes progressively smaller as the frequency is indefinitely increased.

Memory Requirements in a Telephone Exchange

By CLAUDE E. SHANNON

(*Manuscript Received Dec. 7, 1949*)

1. Introduction

A GENERAL telephone exchange with N subscribers is indicated schematically in Fig. 1. The basic function of an exchange is that of setting up a connection between any pair of subscribers. In operation the exchange must "remember," in some form, which subscribers are connected together until the corresponding calls are completed. This requires a certain amount of internal memory, depending on the number of subscribers, the maximum calling rate, etc. A number of relations will be derived based on these considerations which give the minimum possible number of relays, crossbar switches or other elements necessary to perform this memory function. Comparison of any proposed design with the minimum requirements obtained from the relations gives a measure of the efficiency in memory utilization of the design.

Memory in a physical system is represented by the existence of stable internal states of the system. A relay can be supplied with a holding connection so that the armature will stay in either the operated or unoperated positions indefinitely, depending on its initial position. It has, then, two stable states. A set of N relays has 2^N possible sets of positions for the armatures and can be connected in such a way that these are all stable. The total number of states might be used as a measure of the memory in a system, but it is more convenient to work with the logarithm of this number. The chief reason for this is that the amount of memory is then proportional to the number of elements involved. With N relays the amount of memory is then $M = \log 2^N = N \log 2$. If the logarithmic base is two, then $\log_2 2 = 1$ and $M = N$. The resulting units may be called binary digits, or more shortly, bits. A device with M bits of memory can retain M different "yes's" or "no's" or M different 0's or 1's. The logarithmic base 10 is also useful in some cases. The resulting units of memory will then be called decimal digits. A relay has a memory capacity of .301 decimal digits. A 10×10 crossbar switch has 100 points. If each of these points could be operated independently of the others, the total memory capacity would be 100 bits or 30.1 decimal digits. As ordinarily used, however, only one point in a vertical can be closed. With this restriction the capacity is one decimal digit for each vertical, or a total of ten decimal digits. The panels used in a

panel type exchange are another form of memory device. If the commutator in a panel has 500 possible levels, it has a memory capacity of log 500;8.97 bits or 2.7 decimal digits. Finally, in a step-by-step system, 100-point selector switches are used. These have a memory of two decimal digits.

Frequently the actual available memory in a group of relays or other devices is less than the sum of the individual memories because of artificial restrictions on the available states. For technical reasons, certain states are made inaccessible—if relay A is operated relay B must be unoperated, etc. In a crossbar it is not desirable to have more than nine points in the same horizontal operated because of the spring loading on the crossarm. Constraints of this type reduce the memory per element and imply that more than the minimum requirements to be derived will be necessary.

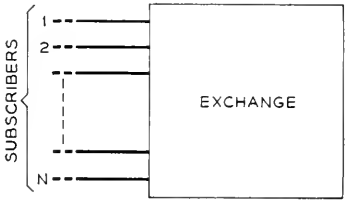

Fig. 1—General telephone exchange.

2. Memory Required for any S Calls out of N Subscribers

The simplest case occurs if we assume an isolated exchange (no trunks to other exchanges) and suppose it should be able to accommodate any possible set of S or fewer calls between pairs of subscribers. If there are a total of N subscribers, the number of ways we can select m pairs is given by

$$\frac{N(N-1)(N-2)\cdots(N-2m+1)}{2^m m!} = \frac{N!}{2^m m!(N-2m)!} \tag{1}$$

The numerator $N(N-1)\cdots(N-2m+1)$ is the number of ways of choosing the $2m$ subscribers involved out of the N. The $m!$ takes care of the permutations in order of the calls and 2^m the inversions of subscribers in pairs. The total number of possibilities is then the sum of this for $m = 0, 1, \cdots, S$; i.e.

$$\sum_{m=0}^{S} \frac{N!}{2^m m!(N-2m)!} \tag{2}$$

The exchange must have a stable internal state corresponding to each of these possibilities and must have, therefore, a memory capacity M where

$$M = \log \sum_{0}^{S} \frac{N!}{2^m m!(N-2m)!}. \tag{3}$$

If the exchange were constructed using only relays it must contain at least $\log_2 \sum N!/2^m m!(N - 2m)!$ relays. If 10×10 point crossbars are used in the normal fashion it must contain at least $\frac{1}{10} \log_{10} \sum N!/2^m m!(N - 2m)!$ of these, etc. If fewer are used there are not enough stable configurations of connections available to distinguish all the possible desired interconnections. With $N = 10{,}000$, and a peak load of say 1000 simultaneous conversations $M = 16{,}637$ bits, and at least this many relays or 502 10×10 crossbars would be necessary. Incidentally, for numbers N and S of this magnitude only the term $m = S$ is significant in (3).

The memory computed above is that required only for the basic function of remembering who is talking to whom until the conversation is completed. Supervision and control functions have been ignored. One particular supervisory function is easily taken into account. The call should be charged to

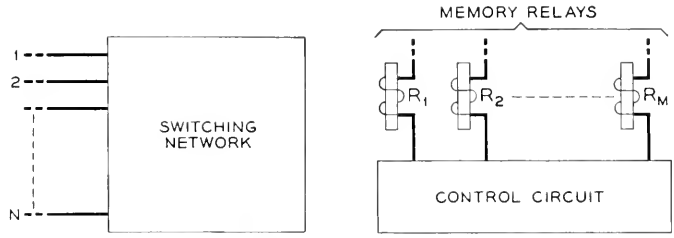

Fig. 2—Minimum memory exchange.

the calling party and under his control (i.e. the connection is broken when the calling party hangs up). Thus the exchange must distinguish between *a* calling *b* and *b* calling *a*. Rather than count the number of pairs possible we should count the number of ordered pairs. The effect of this is merely to eliminate the 2^m in the above formulas.

The question arises as to whether these limits are the best possible—could we design an exchange using only this minimal number of relays, for example? The answer is that such a design is possible in principle, but for various reasons quite impractical with ordinary types of relays or switching elements. Figure 2 indicates schematically such an exchange. There are M memory relays numbered 1, 2, ..., M. Each possible configuration of calls is given a binary number from 0 to 2^M and associated with the corresponding configuration of the relay positions. We have just enough such positions to accommodate all desired interconnections of subscribers.

The switching network is a network of contacts on the memory relays such that when they are in a particular position the correct lines are connected together according to the correspondence decided upon. The control circuit is essentially merely a function table and requires, therefore, no memory. When a call is completed or a new call originated the desired con-

figuration of the holding relays is compared with the present configuration and voltages applied to or eliminated from all relays that should be changed.

Needless to say, an exchange of this type, although using the minimum memory, has many disadvantages, as often occurs when we minimize a design for one parameter without regard to other important characteristics. In particular in Fig. 2 the following may be noted: (1) Each of the memory relays must carry an enormous number of contacts. (2) At each new call or completion of an old call a large fraction of the memory relays must change position, resulting in short relay life and interfering transients in the conversations. (3) Failure of one of the memory relays would put the exchange completely out of commission.

3. The Separate Memory Condition

The impracticality of an exchange with the absolute minimum memory suggests that we investigate the memory requirements with more realistic assumptions. In particular, let us assume that in operation a separate part of the memory can be assigned to each call in progress. The completion of a current call or the origination of a new call will not disturb the state of the memory elements associated with any call in progress. This assumption is reasonably well satisfied by standard types of exchanges, and is very natural to avoid the difficulties (2) and (3) occurring in an absolute minimal design.

If the exchange is to accommodate S simultaneous conversations there must be at least S separate memories. Furthermore, if there are only this number, each[1] of these must have a capacity $\log \frac{N(N-1)}{2}$. To see this, suppose all other calls are completed except the one in a particular memory. The state of the entire exchange is then specified by the state of this particular memory. The call registered here can be between any pair of the N subscribers, giving a total of $N(N-1)/2$ possibilities. Each of these must correspond to a different state of the particular memory under consideration, and hence it has a capacity of least $\log N(N-1)/2$.

The total memory required is then

$$M = S \log \frac{N(N-1)}{2}. \tag{4}$$

If the exchange must remember which subscriber of a pair originated the call we obtain

$$M = S \log N(N-1). \tag{5}$$

or, very closely when N is large,

$$M = 2S \log N. \tag{6}$$

[1] B. D. Holbrook has pointed out that by using more than S memories, each can have for certain ratios of $\frac{S}{N}$, a smaller memory, resulting in a net saving. This only occurs, however, with unrealistically high calling rates.

The approximation in replacing (5) by (6), of the order of $\frac{S}{N} \log e$, is equivalent to the memory required to allow connections to be set up from a subscriber to himself. With $N = 10{,}000$, $S = 1{,}000$, we obtain $M = 26{,}600$

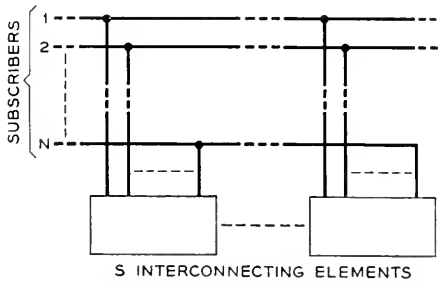

Fig. 3—Minimum separate memory exchange.

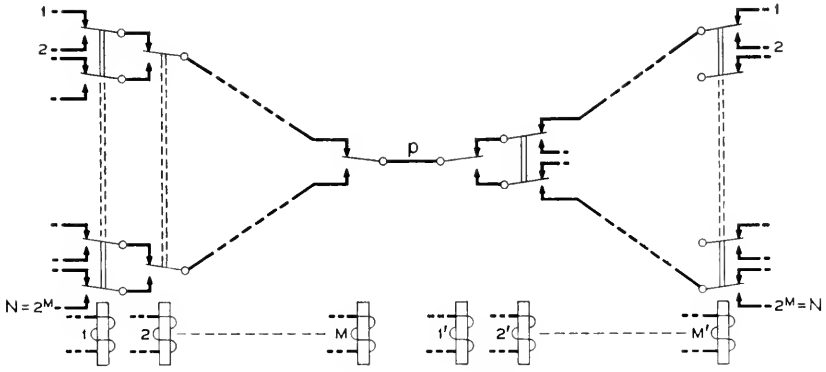

Fig. 4—Interconnecting network for Fig. 3.

from (6). The considerable discrepancy between this minimum required memory and the amount actually used in standard exchanges is due in part to the many control and supervision functions which we have ignored, and in part to statistical margins provided because of the limited access property.

The lower bound given by (6) is essentially realized with the schematic exchange of Fig. 3. Each box contains a memory $2 \log N$ and a contact network capable of interconnecting any pair of inputs, an ordered pair being associated with each possible state of the memory. Figure 4 shows such an interconnection network. By proper excitation of the memory relays 1, 2, $\cdots$, M, the point p can be connected to any of the $N = 2^m$ subscribers on the left. The relays $1'$, $2'$, $\cdots$, M' connect p to the called subscriber on

the right. The general scheme of Fig. 3 is not too far from standard methods, although the contact load on the memory elements is still impractical. In actual panel, crossbar and step-by-step systems the equivalents of the memory boxes are given limited access to the lines in order to reduce the contact loads. This reduces the flexibility of interconnection, but only by a small amount on a statistical basis.

4. Relation to Information Theory

The formula $M = 2S \log N$ can be interpreted in terms of information theory.[2] When a subscriber picks up his telephone preparatory to making a call, he in effect singles out one line from the set of N, and if we regard all subscribers as equally likely to originate a call, the corresponding amount of information is $\log N$. When he dials the desired number there is a second choice from N possibilities and the total amount of information associated with the origin and destination of the call is $2 \log N$. With S possible simultaneous calls the exchange must remember $2S \log N$ units of information.

The reason we obtain the "separate memory" formula rather than the absolute minimum memory by this argument is that we have overestimated the information produced in specifying the call. Actually the originating subscribers must be one of those not already engaged, and is therefore in general a choice from less than N. Similarly the called party cannot be engaged; if the called line is busy the call cannot be set up and requires no memory of the type considered here. When these factors are taken into account the absolute minimum formula is obtained. The separate memory condition is essentially equivalent to assuming the exchange makes no use of information it already has in the form of current calls in remembering the next call.

Calculating the information on the assumption that subscribers are equally likely to originate a call, and are equally likely to call any number, corresponds to the maximum possible information or "entropy" in communication theory. If we assume instead, as is actually the case, that certain interconnections have a high *a priori* probability, with others relatively small, it is possible to make a certain statistical saving in memory.

This possibility is already exploited to a limited extent. Suppose we have two nearby communities. If a call originates in either community, the probability that the called subscriber will be in the same community is much greater than that of his being in the other. Thus, each of the exchanges can be designed to service its local traffic and a small number of intercommunity calls. This results in a saving of memory. If each exchange has N subscribers and we consider, as a limiting case, no traffic between exchanges,

[2] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27, pp. 379–423, and 623–656, July and October 1948.

the total memory by (6) would be $4S \log N$, while with all $2N$ subscribers in the same exchange $4S \log 2N$ would be required.

The saving just discussed is possible because of a group effect. There are also statistics involving the calling habits of individual subscribers. A typical subscriber may make ninety per cent of his calls to a particular small number of individuals with the remaining ten per cent perhaps distributed randomly among the other subscribers. This effect can also be used to reduce memory requirements, although paper designs incorporating this feature appear too complicated to be practical.

ACKNOWLEDGMENT

The writer is indebted to C. A. Lovell and B. D. Holbrook for some suggestions incorporated in the paper.

Matter, A Mode of Motion

By R. V. L. HARTLEY

(*Manuscript Received Feb. 28, 1950*)

Both the relativistic and wave mechanical properties of particles appear to be consistent with a picture in which particles are represented by localized oscillatory disturbances in a mechanical ether of the MacCullagh-Kelvin type. Gyrostatic forces impart to such a medium an elasticity to rotation, such that, for very small velocities, its approximate equations are identical with those of Maxwell for free space. The important results, however, follow from the inherent non-linearity of the complete equations and the time dependence of the elasticity associated with finite displacements. These lead to reflections which permit of a wave of finite energy remaining localized. Because of the non-linearity, the amplitude and energy of a stable mode, as well as the frequency, are determined by the constants of the medium. Such a stable mode is capable of translational motion and so is suitable to represent a particle. The mass assigned to it is derived from its energy by the relativity relation. While this mass is dimensionally the same as that of the medium it is differently related to the energy and so need not conform to the classical laws which the latter is assumed to obey.

Exchanges of energy between particles and between a particle and radiation involve frequency changes as in the quantum theory. The experimental detection of a uniform velocity relative to the medium is not to be expected. Besides providing a new approach to the problems of particle mechanics, the theory offers the prospect of incorporating the present pictures into a more comprehensive one, with a material reduction in the number and complexity of the independent assumptions.

Introduction

THE following quotation states a conclusion which is widely held: "But in view of the more recent development of electrodynamics and optics it became more and more evident that classical mechanics affords an insufficient foundation for the physical description of all natural phenomena."[1] This implies that classical mechanics and classical electromagnetics are so alike that one may be condemned for the shortcomings of the other. Actually, classical electromagnetics is in open disagreement with classical mechanics particularly with respect to those features for which it has been most criticized. According to the mechanical principle of relativity,[2] the equations of any mechanical system are invariant under the Newtonian transformation, $x = x' + Vt', y = y', z = z', t = t'$, where V is a constant velocity in the x direction. Since the classical electromagnetic equations are not invariant under this transformation, they cannot describe the performance of any classical mechanical system. Their failures, therefore, should not stand in the way of a study of the possibilities of such systems.

The system considered here is the so-called rotational ether, suggested

[1] A. Einstein, The Theory of Relativity, Methuen & Co., Ltd., London, 1921, p. 13.
[2] Haas, Introduction to Theoretical Physics, 2nd Ed., Vol. I, p. 46.

by MacCullagh and elaborated by Kelvin, in which the stiffness is associated with gyrostatic forces. Some consideration has been given to an alternative model consisting of a non-viscous liquid in a high state of fine scale turbulence. It is well known that, by virtue of the gyrostatic forces associated with it, a vortex will transmit a wave of transverse displacement along its axis. It would appear, therefore, that a gross wave involving similar displacements would be passed along from vortex to vortex, much as a sound wave is passed from molecule to molecule. However, since this model has not yet been shown to be fully equivalent to Kelvin's, attention will be confined to the latter. While this, as developed by Kelvin, gave a satisfactory description of electromagnetic waves in free space, it had nothing to represent matter. This was assumed to be something different from ether, which might or might not be pervaded by it. A closer study of the model has indicated that the peculiar nature of its stiffness makes possible sustained oscillatory disturbances in which the energy remains localized about a center which may move with any velocity less than that of a free wave. It is proposed to use such quasi-standing wave patterns to describe material particles. Matter, then, has no existence apart from the ether, and the motion of particles is the motion of patterns of mechanical wave motion. While the ether itself conforms to Newtonian mechanics, the mechanics of such a wave pattern, considered as a particle located at its center, is much more complicated than that of the familiar mass point of particle dynamics. This complexity provides a bridge from the older concepts of particle behavior to the new.

The study of this model given below reveals no insuperable obstacles such as were encountered by the electromagnetic theory and the simpler ether model. The properties of the wave-patterns are qualitatively consistent with many of the concepts of modern physics, though in some cases not with the generality of application which is now assigned to them. Among these concepts are: the space-time of special relativity, relativistic mechanics, de Broglie waves, proportionality of energy and frequency, energy thresholds, and transfers of energy according to the quantum frequency formula. The ether model also leads to certain concepts not found in the present theories. It provides, for example, for a possible failure of the mass-energy balance such as has been observed in nuclear reactions. It also suggests the possibility of a new type of particle which, by virtue of its negative inertial mass, is capable of exerting a binding force between other particles.

These results make it more probable that classical mechanics may, after all, afford a sufficient "*foundation* for the physical description of all natural phenomena" even though the super-structure be very different from that contemplated by its originators. The present argument, however, is not that this particular description is necessary, but rather that it offers distinct

advantages. On the philosophical side, there is the prospect of greater unification of the basic theory through a reduction in the number of independent assumptions. Matter and radiation appear as wave motions which satisfy the same equations. The apparent conflicts between current concepts appear to be reconcilable through a more exact determination of the conditions under which each applies. On the more practical side, the ether model provides a different approach and technique. It has the advantage inherent in all models that, once one is found which fits one set of conditions, a study of its properties under widely different conditions may bring out relations which it would be difficult to postulate solely on the basis of observations made under the second conditions. The suggested existence of particles having negative inertia, as discussed near the end of the paper, should it lead to anything of value, would be an example of such a relation. Also it makes available the added relationships which are characteristic of non-linear equations, without encountering those difficulties with respect to absolute motion which may arise when non-linearity is introduced arbitrarily. While the working out of the quantitative relations involved is a rather formidable undertaking, any effort in that direction may well throw new light on those problems which have not yielded to other methods.

The Gyrostatic Ether

As stated above the specific form of gyrostatic medium on which the present discussion is based is the ether model proposed by Kelvin. This is discussed in detail in a companion paper.[3] It is there shown that, for infinitesimal displacements, it is characterized by the wave equations:

$$\nabla \times \left(\frac{\overline{T}}{2}\right) = \rho_0 \frac{\partial \bar{q}}{\partial t} \tag{1}$$

$$\nabla \times \bar{q} = -\frac{1}{\eta_0} \frac{\partial}{\partial t}\left(\frac{\overline{T}}{2}\right), \tag{2}$$

where ρ_0 is the density, η_0 is a generalized stiffness determined by the constants of the medium, $\bar{q}$ is the vector velocity, and $\overline{T}$ is a vector torque per unit volume, which has its origin in the torque with which a gyrostat opposes an angular displacement of its axis. For a plane polarized plane wave, the quantity $\frac{\overline{T}}{2}$ can be interpreted as a surface tractive force per unit area, which a layer of the medium normal to the direction of propagation exerts on the layer just ahead. Its direction lies in the surface of separation, and is parallel to that of the velocity $\bar{q}$.

[3] R. V. L. Hartley, "The Reflection of Diverging Waves by a Gyrostatic Medium"—this issue of *The Bell System Technical Journal*.

These equations become identical with those of Maxwell for free space,

$$\nabla \times \bar{H} = \epsilon \frac{\partial \bar{E}}{\partial t},$$

$$\nabla \times \bar{E} = -\mu \frac{\partial \bar{H}}{\partial t},$$

if we replace $\bar{q}$ by $\bar{E}$, $\frac{\overline{T}}{2}$ by $\bar{H}$, ρ_0 by ϵ and $\frac{1}{\eta}$ by μ. Then $\rho_0 \bar{q}$ corresponds to $\bar{D}$ and -2φ to $\bar{B}$ where φ is the angular displacement of an element of the medium. Or the roles of the electric and magnetic quantities may be interchanged.

For present purposes, however, we are more interested in finite displacements. The relations which then apply are discussed in detail in the companion proper. It is there shown that changes of two kinds appear in (1) and (2), with corresponding changes in the transmission properties of the medium. The simple linear relations are to be replaced by non-linear ones, which cause distortion of a wave but no reflection. In addition, a qualitative difference appears in the nature of the elasticity, as was pointed out by Kelvin. The restoring torque is no longer proportional to the angular displacement alone. When the axis of a gyrostat is displaced it begins rotating toward the axis of the displacement, thereby decreasing the component of its spin which is normal to that axis. Thus the restoring torque for a constant angular displacement decreases with time. The restoring torque is therefore a function of the time as well as of the displacement. Because of this time dependence, a disturbance of finite amplitude generates waves which propagate both backward and forward.

For a plane progressive sine wave it is found that the reflected waves interfere destructively. However, if a central generator starts sending out a diverging sinusoidal disturbance, a part of the energy is reflected inward as a wave of the same frequency as the generator and another smaller part as waves the frequencies of which are odd multiples of that frequency. This reflection attenuates the outgoing wave. If the incoming wave is reflected rather than absorbed at the generator, it tends to set up a standing wave pattern. As time goes on, the impedance of the medium as seen from the generator becomes more reactive and less power is drawn from the generator. Due to the attenuation, the energy in spherical shells of a given thickness decreases with increasing radius, so that it and the power transmitted at the wave front approach zero as r approaches infinity. This falling off is somewhat similar to that suffered by a wave the frequency of which lies in the stop band of a filter, but with one important difference. There the attenuation is independent of the distance. But here, since the attenuation is a

function of the magnitude of the disturbance and of the curvature of the wave-front, the attenuation constant approaches zero as r increases indefinitely.

Whether or not the total energy stored in the wave pattern will approach a finite or infinite value depends on how fast the attenuation decreases with distance, and a more complete solution is needed to give an exact answer. If it does approach infinity it will do so much more slowly than for a medium which does not reflect.

The disagreement between classical electromagnetics and mechanics, referred to above, may now be stated more explicitly. The former says that electromagnetic waves are represented exactly by Maxwell's equations, regardless of the magnitudes of the electromagnetic variables. When these waves are interpreted as existing in a mechanical ether, classical mechanics says that Maxwell's relationship is approached as a limit as the mganitudes approach zero. Waves of finite amplitude are to be represented by the more complicated relations.

The two systems differ in three important respects; their relation to uniform linear motion, the linearity of their equations and the nature of the elasticity involved. Because the classical electromagnetic equations are not invariant under a Newtonian transformation, the set of axes to which the equations refer are uniquely related to other sets which are moving uniformly with respect to them. In special relativity, this condition is avoided by modifying the classical concepts of space and time to conform to the fact that the equations are invariant under the Lorentz transformation. The Newtonian invariance of the ether equations, however, insures that a set of axes at rest with respect to the undisturbed ether is not unique. Hence in the modified model, in which *the motions which constitute matter* conform to the laws of the ether, a uniform linear velocity of the entire system cannot be detected. This is consistent with the accepted principle that absolute velocity is meaningless.

We are, however, still faced with the question of the detection of uniform motion of matter relative to the ether. This is discussed at length below, where it is shown that the properties of the ether lead directly to an auxiliary space-time, which applies very closely under the experimental conditions and accounts for the failure to detect the motion. This "experimental" space-time is formally identical with that of special relativity. Thus the modification of the space-time of classical electromagnetics which appears in special relativity might be said to bring it into closer formal agreement with the classical mechanics of ether wave patterns. At any rate the establishing of this theoretical connection between the space-time of special relativity and a classical mechanical model is a step toward unification.

On the matter of linearity, proposals have been made to add arbitrary non-

linear terms to Maxwell's equations. While this also makes the electromagnetic equations more like those of the ether, an important difference still remains. An equation obtained in this way is not necessarily invariant under either a Newtonian or a Lorentz transformation. If, then, the axes with respect to which it is expressed are not to be unique, it must be shown that some transformation exists under which it is invariant. Not only is the form of the equation important here but also the interpretation of the dependent variables. For example, since the complete equations of the ether contain $q\cdot\nabla$, if the mechanical variables be replaced by the analogous electromagnetic ones, the equations will be Newtonian invariant only if $\dot{E}$, which replaces $\dot{q}$, is interpreted as a velocity. It is evident, therefore, that the fact that we are dealing with a mechanical model is an important point in the argument. Also, unless the added terms make the effective constants depend on the time as well as the dependent variables, there will be no reflection of the energy in a finite disturbance and the medium will not have the energy trapping property which is essential to the present argument.

Stationary Wave Patterns

The first question to be considered is the possibility of setting up a sustained wave pattern suitable to represent a particle at rest with respect to the ether. The simplest procedure might seem to be to look for it as a solution of the approximate linear equations in the form of a pair of spherical waves propagating radially, one outward and one inward, so as to form together a standing wave pattern. However, certain difficulties are encountered. There is nothing in the free linear ether which can serve as boundary conditions to fix the position or size of the pattern. Even if these were determined, there would be nothing to fix the amplitude, and so the energy. Most patterns, particularly those which involve a single frequency, have one or more of the following features. Some of the variables become infinite at the center; the total energy is infinite, energy is propagated away radially.

These difficulties disappear, however, when we take account of the properties of the ether for disturbances of finite amplitude. Let us suppose that the energy which is to constitute the pattern is supplied by a central generator, the impedance of which is mainly reactive, so that reflected waves which reach it are reflected outward again. Once a standing wave pattern has been established as described above, let the force of the generator be reduced to zero without changing its impedance. The pattern will then persist except for a small and decreasing damping due to the outward radiation at its periphery. However, in the region near the center the displacements will be very large, and the incoming reflected waves will suffer reflec-

tions which increase with decreasing radius. These reflections will effectively take the place of the assumed reactive impedance of the generator, and so the latter may be discarded. The fact that the reflections take place from a somewhat diffuse inner boundary prevents the amplitude from building up to an infinite value at the center as it would with a linear medium.

However, the reflected wave includes components of triple and higher frequencies and, due to the non-linearity, other frequency components will be generated. If the entire pattern is to be stable, all of these must satisfy the boundary conditions. Their magnitudes relative to the fundamental, for a particular mode of oscillation, will depend on the amplitude and frequency of the fundamental, as well as on the constants of the medium. Hence the amplitude as well as the frequency of a stable pattern of a particular mode should be uniquely determined. Particles of different properties would then be expected to consist of patterns involving different modes of oscillation.

Returning to the lack of complete reflection at the outer boundary and the change it might be expected to make in the pattern with time, this might be an important factor for a single particle alone in the universe. Actually, however, a very large number of particles are present. If we consider a point at a considerable distance from any one particle, a point in a vacuum, the resultant of the disturbances produced there by all the patterns will be very large compared with that due to any one. But the effect on a particular pattern of its own loss by radiation will be determined by this small component, and so will be small compared with the effect exerted on it by the combined small fields of its neighbors. This combined field due to a large number of patterns, randomly placed, and moving at random, will constitute a randomly varying electromagnetic field in a vacuum, such as has recently been postulated for other reasons. If, now, the center of a pattern be placed at the point in question, this random field may occasionally take on so large a value as to disturb the equilibrium conditions of the pattern.

It may be argued that, in spite of the merging of a given pattern in that of the random group, the group as a whole will suffer a progressive loss of energy through incomplete reflection. Were this to occur the total loss of energy would not be evenly distributed among the particles. As discussed below the particles would exchange energy through the mechanism of the non-linearities, continually forming less stable group patterns of greater energy, which in turn suffer transitions to more stable patterns of lower energy. A small continuous decrease in total energy would manifest itself as an increase in the rate of transitions downward in energy compared to those upward.

Associated with a standing wave pattern such as that described above

would be three regions. Near the center would be a relatively small core in which the non-linear effects predominate and linear theory is totally inapplicable. Farther out the departure from linearity is only moderate, and the variation of the constants with distance is slow enough that the reflections are small. It should be possible to treat wave propagation in this region by the methods developed for a string of variable density, which are sometimes cited as analogous with those employed in wave mechanics. The analogy is made closer by the fact that the variations in impedance which correspond to the varying density are determined by the energy density of the pattern itself. Still farther out the amplitudes become still smaller, the ether constants become very nearly but not quite uniform, and the pattern approaches very closely to that in a linear medium.

While the nature of the pattern is determined largely by the non-linear inner region, because of the small volume of this region most of the energy will be located in the nearly linear region. So we might expect some at least of the macroscopic properties of the pattern to differ very little from those deduced from a consideration of the corresponding pattern in a linear medium. We will therefore begin by examining such a pattern. For the linear case, when the axes are at rest with respect to the undisturbed ether, (1) and (2) lead to the wave equation for the vector displacement $\bar{s}$,

$$\frac{\partial^2 \bar{s}}{\partial t^2} = c^2 \nabla^2 \bar{s}. \tag{3}$$

As is well known, this is satisfied by any function of the form

$$\bar{s} = f(\omega t \pm k_x x \pm k_y y \pm k_z z),$$

where

$$\frac{\omega^2}{c^2} = k_x^2 + k_y^2 + k_z^2, \tag{4}$$

and the constants ω, k_x, k_y and k_z, are real or complex. Since an imaginary frequency is interpreted as an exponential change with time, it is not suitable for representing a permanent pattern, so ω will be taken to be real. Imaginary values of k are interpreted as exponential variations with distance. But, since $\bar{s}$ is always real, we may, by a four-dimensional Fourier analysis, represent f as the summation of components of the form

$$\bar{s} = \bar{A} \cos (\omega t \pm k_x x \pm k_y y \pm k_z z), \tag{5}$$

where $\bar{A}$ is a complex vector representing the amplitude and phase of the component, and k_x, k_y and k_z are real. Since each component must satisfy (3), the new constants must satisfy (4). Each such component constitutes a plane progressive wave traveling, with velocity c in a direction, the cosines of which are proportional to the wave numbers k_x, etc.

As a first step in building up a stationary pattern, in which there is no steady propagation of energy in any direction, we combine two progressive wave components (5) which are identical, except that their directions of phase propagation along, say, the z axis are opposite. The signs of the last terms are then opposite and the sum can be written

$$\bar{s} = 2\bar{A} \cos (\omega t \pm k_x x \pm k_y y) \cos k_z z.$$

Proceeding in the same way for x and y, we arrive at the standing wave pattern,

$$\bar{s} = 8\bar{A} \cos \omega t \cos k_x x \cos k_y y \cos k_z z. \tag{6}$$

Components of this sort, each with its own amplitude and phase, may be combined to build up possible stationary patterns. However, we shall not attempt here to build such patterns, but rather to deduce what information we can from a study of a single component.

Moving Wave Patterns

In order to represent approximately a particle in uniform linear motion, we are to look for a solution of (3) which represents a moving wave pattern. For this we make use of two functions which may readily be shown to be such solutions,

$$\bar{s} = g_+ \left(\beta(\omega + Vk_x)t - \beta \left(k_x + \frac{V\omega}{c^2} \right) x \pm k_y y \pm k_z z \right),$$

$$\bar{s} = g_- \left(\beta(\omega - Vk_x)t + \beta \left(k_x - \frac{V\omega}{c^2} \right) x \pm k_y y \pm k_z z \right),$$

where ω, k_x, k_y and k_z are real and satisfy (4), V is a real constant, and

$$\beta^2 = \frac{1}{1 - \frac{V^2}{c^2}}.$$

g_+ represents a plane progressive wave the propagation of which along the x axis is in the positive direction. g_- represents one of lower frequency, propagating in the negative x direction. Their wave numbers in the x direction differ in such a way that those in the y and z direction are the same for the two. In the plane wave case, where $k_y = k_z = 0$ and $\omega = ck_x$, they reduce to

$$\bar{s} = g_\pm \left(\beta \left(1 \pm \frac{V}{c} \right) \omega \left(t \mp \frac{x}{c} \right) \right).$$

The two waves then travel in the x direction with velocities c and $-c$, and their frequencies are in the ratio $\dfrac{c + V}{c - V}$.

In order to derive a quasi stationary pattern we replace the functions $g_+(\ )$ and $g_-(\ )$ by $\bar{B}\cos\alpha(\ )$ and combine components in a manner similar to that used in deriving (6). The result is

$$\bar{s} = 8\bar{B}\cos\alpha\beta\omega\left(t - \frac{V}{c^2}x\right)\cos\alpha\beta k_x(x - Vt)\cos\alpha k_y y\cos\alpha k_z z, \quad (7)$$

where $\bar{B}$ is a complex vector, and α may be any real scalar function of V. When we compare this with (6) we find that the last three factors, which in (6) describe a fixed envelope, in (7) describe an envelope which moves in the x direction with velocity V. For the same values of k_x, k_y and k_z, the moving pattern has its dimensions in the x direction reduced relative to those in the y and z in the ratio $\frac{1}{\beta}$. The first factor in (6) describes a sinusoidal variation with time which is everywhere in the same phase. In (7) it describes one, the phase of which varies linearly with x. This factor also describes a wave which progresses in the x direction with a velocity $\frac{c^2}{V}$. The existence of such a wave as a factor in the expression for a moving wave pattern was commented on by Larmor.[4] Aside from the constant α in (7) it will be recognized as the Lorentz transform of (6), as it should be since the approximate equations of which it is a solution are invariant under this transformation.

We shall take (7) to represent one component of a moving wave pattern which represents a moving particle. If we transform this to axes moving with the pattern by a Newtonian transformation it becomes

$$\bar{s} = 8\bar{B}\cos\alpha\left(\frac{\omega}{\beta}t' - \frac{\beta\omega V}{c^2}x'\right)\cos\alpha\beta k_x x'\cos\alpha k_y y'\cos\alpha k_z z', \quad (8)$$

in which the envelope is at rest. This may be thought of as a stationary wave in an ether which is moving relative to the axes with a velocity $-V$. It is a solution of the wave equation for such an ether, as obtained by transforming (3) to the moving axes, or

$$\frac{\partial^2\bar{s}}{\partial t'^2} = c^2\nabla'^2\bar{s} + 2V\frac{\partial^2\bar{s}}{\partial x'\,\partial t'} - V^2\frac{\partial^2\bar{s}}{\partial x'^2}.$$

The one dimensional form of this equation is identical with that given by Trimmer[5] for compressional waves in moving air, except that in one case $\bar{s}$ is solenoidal and in the other divergent.

So far we have found no reason to associate any particular moving pattern with the assumed stationary one, in the sense that the moving pat-

[4] Larmor, Ency. Brit. 11th Ed., 1910; 13th Ed., 1926, Vol. 22, p. 787.
[5] J. D. Trimmer, *Jour. Acous. Soc. Am.*, 9, p. 162, 1937.

tern describes the result of setting in motion the particle which is described by the stationary pattern. Without further knowledge or assumptions regarding the factors which control the form of the pattern, we can go no farther in this direction by theory alone. Rather than try to guess at these factors, it seems preferable to investigate what properties the wave patterns must have in order to conform to the known results of experiment.

Let us start with the Michelson-Morley experiment to which the earlier ether theory did not conform. The entire apparatus involved in the experiment is now to be considered as made up of particles each of which consists of a wave pattern in the ether. The apparatus as a whole may be regarded as a more complicated wave pattern. The interference pattern formed by the light beams may, if we wish, be included in the over-all pattern. The results to be expected in the experiment do not depend on the oscillatory nature of the wave, nor on its amplitude or phase, but only on its spatial distribution, which is determined by the envelope factors. It is obvious from (8) that, for any uniform velocity $-V$ of the ether relative to the apparatus, the ratios of the dimensions of the envelope along the motion to those across it are reduced, relative to their values when V is zero, in the ratio $\frac{1}{\beta}$. That is to say the apparatus like the fringes undergo this change in relative dimensions. But, as is well known, this is exactly what is required in order that there shall be no apparent motion of the fringes. Hence any one of the stationary patterns in a moving ether, as represented by (8), is consistent with the experiment. This experiment therefore furnishes no basis for selecting any particular pattern.

More generally, in any experiment, the distances and time intervals which are available as standards of comparison are associated with the wave patterns and change with their motion. Thus we may, following the special theory of relativity, define an auxiliary space and time, the units of which are associated with the dimensions and cyclic interval of a particular periodic wave pattern. This pattern then plays the roles of the "practically rigid body" and the "clock" which determine space and time in relativity theory. An examination of (8) shows that the dimensions of the pattern, its frequency, and its phase change with the velocity of the ether relative to the pattern in just the way that the corresponding quantities associated with the rigid body and clock change with velocity in the relativity theory. But there these changes are known to be such that no experiment can detect the velocity involved. It follows, therefore, that no experiment in which the apparatus consists of wave patterns of small amplitude is capable of detecting the velocity V, in (8), which in this case is the velocity of the ether relative to the apparatus. Hence any of the above patterns are consistent with the failure of all experiments designed to detect motion

relative to the ether. When account is taken of the non-linearity of the ether the result to be expected should differ from that just found for the linear case only by the small difference between the linear and non-linear patterns, which may easily be too small to measure. Thus the principal obstacle to the older ether theory is removed.

While the special theory of relativity is usually written in the form which corresponds to α being unity in (8), it has long been recognized that there is no theoretical basis for this particular value. The ether patterns are consistent with the more general formulation. In order to pin down the value of α for the ether patterns we resort to another experiment. Ives and Stillwell[6] found that a molecule which emits radiation of frequency ω when at rest emits a frequency $\frac{\omega}{\beta}$ when in motion. This moving frequency is taken relative to axes moving with the molecule, and so is to be compared with the frequency of oscillation $\frac{\alpha}{\beta}\omega$ in (8). This indicates that in order to represent a component of the pattern which results when the fixed pattern is set in motion, we are to put α equal to unity.

Another observed relation is that the energy of a moving particle is β times that of the same particle at rest. This information should be useful in checking any theory of the mechanism by which the non-linearity of the medium determines the energy of the pattern. All we shall do here is to point out one relation, the significance of which from the standpoint of mechanism will be discussed below. In (7), where the frequency is expressed relative to the same axes as the energy of the moving pattern, if we put α equal to unity, the frequency also varies as β. Hence if the pattern conforms to experiment with respect to its energy, the energy must be proportional to the frequency.

Obviously, if we define the mass of the particle-pattern as its energy over c^2, the particle will conform to relativistic mechanics. The mass of a particle as so defined, while dimensionally the same as that of the ether, is in other respects quite different. Since it is derived from the energy associated with a disturbance of the ether, it would be zero in the undisturbed ether, while the ether mass would be finite. The momentum of a particle would be determined by the flow of energy associated with it. Also within a particle, if the mode of oscillation were such that the wave propagated continuously around the axis in one direction, the resulting rotation of the energy would be interpreted as an angular momentum or spin. This concept of spin was suggested by Japolsky[7] in connection with cylindrical waves in a linear medium. There is, therefore, no *a priori* reason to expect that the motion

[6] H. E. Ives and C. R. Stillwell, *Jour. Opt. Soc. Am.*, 28, 215, 1938 and 31, 369, 1941.
[7] N. S. Japolsky, *Phil. Mag.* 20, 417, 1935.

of particles should conform to the laws of classical mechanics. As just noted, it should conform much more closely to those of relativistic mechanics. Also, to the extent that the flow of energy follows the laws of wave mechanics, as suggested below, the behavior of the particles will also conform to those laws. Similar considerations apply to the mass of radiation as derived from its energy.

Another experiment which helps to fix the required properties of the patterns is that of Davisson and Germer, in which it is shown that a particle moving with velocity V is diffracted as if it had a wave length λ such that

$$\lambda = \frac{h}{\beta m_0 V},$$

where h is Planck's constant and m_0 is the rest mass.

If, in (7) with α unity, we assume the energy frequency ratio to be equal to h, the wavelength associated with the first factor reduces to the value given by experiment. This does not mean that an ordinary physical wave of this length is present in the pattern. It does mean that, at any instant, the amplitude of the sinusoidal variation of displacement with distance, as given by the remaining factors, varies sinusoidally with the wave length λ, and is zero as points separated by $\frac{\lambda}{2}$. Hence, when the presence of equally spaced obstacles calls for zero values of displacement at equally spaced intervals, the distorted wave should be capable of forming a stable diffraction pattern when the translational velocity of the pattern is such that the interval between points of zero displacement has the value required by the spacing of the obstacles.

Thus the wave pattern will conform to this experiment provided, first, that it is characterized by a particular wave length, and second, that the factor of proportionality between its energy and frequency is equal to h. The first requirement implies that the wave pattern when at rest has practically all of its energy associated with components which are all of the same frequency, or else are confined to a narrow band near the characteristic frequency.

At this point let us pause for a short review and discussion. Briefly, we have replaced the "rigid body" of special relativity by an oscillatory motion of the ether, the envelope of which is analogous with the configuration of the rigid body. We have found that when in motion this envelope behaves as does the rigid body, and the time relations conform to those of a moving clock. These latter may also be interpreted as a multiplying factor which has the form of a plane wave of the DeBroglie type. In wave mechanics, this is treated as a wave of a single frequency and of a variable phase velocity greater than that of light. In the ether theory this wave is interpreted

as one factor in the description of an interference pattern which results from the superposition of component progressive waves of different frequencies, each of which travels with velocity c. This difference in viewpoint leads to other differences.

One of these has to do with the possibility of describing accurately both the position and velocity of a particle, which is ruled out from the wave mechanics viewpoint. An ether wave pattern, however, may have its position accurately described by its envelope, while at the same time the pattern moves with a definite velocity. The particle velocity may here be regarded as a group velocity derived from two waves progressing in opposite directions, but does not depend on the presence of dispersion as does that for waves in the same direction. It is not to be concluded from this that the position and velocity can be *measured* with this accuracy, for we have still to deal with the disturbing effect of the measurement.

From the ether viewpoint, one of the limitations of wave mechanics is to be expected, its inability to calculate directly the position of a particle. The information regarding this position is contained in the expression for the envelope, while the wave factor depends only on its state of motion. A calculation based on a solution which involves the wave factor without the envelope would be expected to be indefinite regarding position. We should expect, however, that it would give information as to the probability of the presence of the particle in a given region, since this is derivable from its state of motion.

Returning to the comparison with experiment, while wave patterns based on the linear equations have shown close agreement so far, the next experiment upsets the applecart. It has been observed that the motion of one particle is modified by the presence of other particles in its neighborhood. So long as the assumed equations are linear, the law of superposition holds, and every solution is independent of every other one. So any wave pattern, when once set up, will continue in its state of rest or of uniform motion indefinitely, and will not be influenced by the presence of other patterns or of free progressive waves. But these together comprise all other matter and radiation. Hence, while we have provided for the property of inertia, there is nothing which tends to alter the state of motion of a body, that is, there are no forces. In this respect the present linear treatment is similar to the special theory of relativity. So, in order to represent the interactions between particles, account must be taken of those between patterns which result from the non-linearity and time dependence of the ether.

Reactions between Patterns

The general problem of the effect of one pattern on another is even more intricate than that of the stable state of a single pattern, which it includes,

and its solution will not be attempted here. Some conclusions may, however, be drawn. Since the amount of reflected energy generated by an element of the medium depends on powers of the instantaneous disturbance higher than the first, the superposition of a second pattern will alter the standing wave pattern of the first, and vice versa. Also, as pointed out in the companion paper, the propagation of both the main and reflected waves also depends on higher powers of the instantaneous disturbance there. The resulting variations in the propagation will also affect the conditions for a stable pattern. Neither pattern, then, can satisfy its stability conditions independently of the other; but if the combined patterns are to be stable they must together satisfy a new set of conditions common to both. How much each is altered by such a union will depend on the degree of coupling between them, that is, on the amount of energy which must be regarded as mutual to the two.

The effect of this coupling will be very different, depending on whether the frequencies of the two patterns are the same or different. When they are different the non-linear terms give rise to frequencies related to the first two by the quatum formula. The transfer of energy to these frequencies may, under favorable conditions, set up a new mode of oscillation the stability conditions of which are better satisfied than those of the original frequencies. The new mode might be that of an excited atom. Or the frequency of one or both of the patterns may be changed to that corresponding to the particle in motion with a particular velocity. In either of these processes some of the energy may be released as radiation at one of the difference frequencies.

If, however, the frequencies of the two patterns are identical, no new frequencies will result from their superposition. If the combined pattern is to persist there must be a stable mode for the combination, the frequency of which is identical with that of the separate patterns. This is hardly to be expected. Also the oscillations of the second pattern, being of the same frequency as those of the first, would have a much greater disturbing effect on its conditions for stability. It would appear, then, that if it were possible to bring two patterns of identical frequency into superposition, they would mutually disintegrate. This does not mean that two particles of the same type cannot exist in the same neighborhood. If they have different velocities, for example, their frequencies will be different. The similarity of these considerations to Pauli's exclusion principle is obvious.

If the second pattern has much greater energy than the first, as it will if it represents a much heavier particle, its stability conditions may be little affected by the presence of the first. The behavior of the first, an electron, may then be discussed on the assumption that it exists in a medium, the properties of which vary with position in accordance with the fixed pattern

of the second particle, the nucleus. Since the stability conditions for the electron pattern particle are most strongly influenced by the effective constants of the medium near its center, we would expect its energy and frequency to be controlled largely by that part of the nuclear pattern which is near its center. Let us assume that, through some external agency, the center of the electron pattern is transferred from one position of rest to another which is differently placed relative to the nucleus. Owing to the different effect of the nuclear pattern on the effective constants of the medium as viewed by the electron pattern, the stable energy of the latter would be different at the second position. This change in rest energy with position may be interpreted as a measure of the change in a field of static potential associated with the massive nucleus. The similarity between this relationship and that which exists between the electron and the nuclear potential in wave mechanics is obvious.

In speaking of a change in the effective constants of the medium, we refer to an average value taken over a number of cycles and wave lengths of the oscillations which make up the second pattern, or nucleus. Calculations based on this concept should not therefore be expected to give valid results when the time intervals involved in the averages are comparable to the period $\frac{h}{m_0c^2}$ of the second particle at rest, or the distances are comparable to the corresponding wave length $\frac{h}{m_0c}$ of the pattern. For a proton this period is 4.38×10^{-24} seconds and the wave length is 1.31×10^{-13} cms. If, then, an electron is to be subject to the kind of nuclear potential field just described, the linear dimensions of that part of it which is controlled by the potential field of the proton must be at least of the order of 10^{-13} cm. This is consistent with Gamow's[8] observation that "It seems, in fact, that a length of the order of magnitude of 10^{-13} centimeters plays a fundamental role in the problem of elementary particles, popping out wherever we try to estimate their physical dimensions."

The variations in the medium due to the nucleus might be treated in terms of their effect on the progressive wave components, the interference of which gives rise to the wave pattern of the electron. The component waves as so influenced should combine to form an interference pattern which represents the behavior of the electron in the field of the nucleus. It is also possible that a technique may be found for treating their effect on that factor of the electron wave which is similar to the DeBroglie wave. This should be more nearly like the techniques now used in wave mechanics.

If two particles are brought so close together that the central cores of their patterns overlap, the departure from linearity becomes so great that

[8] G. Gamow, Physics Today, 2, p. 17, Jan., 1949.

a procedure which may be successful at intermediate separations becomes inadequate. Relativistic mechanics breaks down and Lorentz invariance may lose its significance. This is in agreement with the experimental result that, in some nuclear reactions, the energy balance, as calculated from the relativistic relations, is not satisfied. Also the difficulty which has been encountered in calculating nuclear phenomena by the techniques of wave mechanics suggests that the extremely non-linear condition is approached for the separation of the particles within a nucleus. This viewpoint suggests that an understanding of the nucleus might make possible an experimental determination of velocity relative to the ether.

The reactions between wave patterns of appreciable amplitude may also be viewed from a somewhat different angle. We may think of the various wave patterns as being the analogs of the various modes of motion of, say, an elastic plate. For very small amplitudes they have negligible effect on one another. For larger amplitudes, where Hooke's law does not hold, the force may be represented as a power series of the displacement. The first power term represents the linear stiffness. If the frequencies of two modes which are in oscillation are ω_1 and ω_2, the higher power terms represent forces of frequencies $m\omega_1 \pm n\omega_2$ where m and n are integers or zero. These forces set all the modes into forced oscillation at the frequencies of the various forces, in amounts which depend on the impedance of the particular mode for the particular frequency. When the frequency of the force coincides with the resonant frequency of one of the natural modes, the forced oscillations may be large. Thus the variation in stiffness with displacement provides a coupling whereby energy may be transferred from one or more modes, that is wave patterns, to other modes. But in this transfer the energy always appears associated with a new frequency which is related to those of the modes from which it came in accorance with the familiar formula of quantum theory.

The theory of such energy transformations with change of frequency has been worked out in considerable detail for vacuum tube and other variable resistance modulators, and the results show little in common with the quantum theory beyond the relations connecting the frequencies. When, however, the variation is not in a resistance but in a stiffness, as occurs in the ether case, the situation is quite different. This problem has been explored both theoretically[9] and experimentally.[10] It is found that an oscillation of one frequency in one mode may provide the energy to support sustained oscillations of two other lower frequencies in two other dissipative modes. For this to occur the frequencies involved must be related through the quantum formula. Also the amplitude of the generating oscillation must exceed a

[9] R. V. L. Hartley, *Bell Sys. Tech. Jour.*, 15, 424, 1936.
[10] L. W. Hussey and L. R. Wrathall, *Bell Sys. Tech. Jour.*, 15, 441, 1936.

threshold value which depends on the frequencies, the impedance involved, and the constant of non-linearity. The transformed energy divides itself between the generated modes in the ratio of their frequencies. In a non-dissipative system, the frequencies of possible combinations of sustained oscillations are determined by the energy of the system. Here also they are connected by the quantu n formula.

The particle wave pattern discussed above would approximate very closely to such a non-dissipative non-linear system. We should therefore expect its frequency to be related to its energy through the constants of the ether. In the more complex wave patterns associated with more than one particle, it is unlikely that the pattern representing, say, an electron could maintain its identity as part of some arbitrarily chosen pattern, the magnitudes of which are not commensurable with its own. This suggests that the stable states of the complex pattern would be confined to a sequence of discreet patterns which are related to one another through some property of the electron. These possible non-dissipative combinations of energy and frequency would represent the stable quantum states of the atom. The radiation process would then be similar to that referred to above in which energy from a source of higher frequency distributes itself between two lower frequencies in the ratio of the frequencies. The energy in the pattern of an excited atom would serve as the source. One of the two lower frequencies would be that of a pattern corresponding to a lower energy state to which the transition occurs. The other would be that of the radiating wave which carries off the energy lost in the transition.

A Suggested New Particle

We saw above that the observed variation of the energy of a particle with its velocity calls for a mechanism in which the energy varies directly as the frequency. The fact that a system, in which the stiffness varies with the displacement, is characterized by this relation suggests that the energy of a particle pattern depends mainly on variations in the stiffness of the ether. However, the non-linearities of the ether equations cannot all be interpretated as variable stiffnesses. The non-linearity which appears in (1) when the displacements are finite is equivalent to a variable inertia. It is in order, therefore, to inquire into the properties of a pattern in which the energy is determined by this kind of non-linearity. The variable inductance of an iron-core coil constitutes such a variable inertia. Theoretical and experimental studies of circuits involving these coils have shown that they behave very much as do systems having variable stiffness, with one important exception. The energy distributes itself in the inverse ratio of the frequencies.

If, then, we assume that the energy of a moving pattern is determined by

a mechanism which conforms to this relation, it follows from (7) that its energy will vary as $\frac{1}{\beta}$. Expanding in the usual manner we then have

$$W = m_0c^2 - \tfrac{1}{2}\, m_0V^2 + \cdots$$

This says that a particle represented by such a wave pattern would have a positive rest mass and a negative inertial mass. Its momentum is directed oppositely to its velocity, and energy must be taken from it to set it in motion and given to it to stop it. Such a particle, when bouncing back and forth between two rigid walls or rotating about two centers of force, would exert a force tending to draw them together, instead of the usual repulsion. It is interesting to speculate that if, in an atomic nucleus, the positive charges which are passed back and forth between other nuclear particles were associated with particles of this type their motion would exert a binding force on the other particles.

Conclusion

It appears, then, that the ether model is capable of sustaining wave patterns the behavior of which is qualitatively in agreement with the results of experiment. In order to establish fully the sufficiency of classical mechanics for the physical description of natural phenomena, it will be necessary to work out the complicated quantitative relations whereby the constants of the ether may be deduced from experimental measurements. However, until a serious attempt to do this has failed for some reason other than sheer mathematical complexity, the insufficiency of classical mechanics can scarcely be argued.

In conclusion, I wish to acknowledge the contributions of those of my colleagues who, through discussions over the years, have helped in developing the concepts which have been put together in the above picture.

The Reflection of Diverging Waves by a Gyrostatic Medium

By R. V. L. HARTLEY

(*Manuscript Received Feb. 28, 1950*)

This paper furnishes the basis for a companion one, which discusses the possibility of describing material particles as localized oscillatory disturbances in a mechanical medium. If a medium is to support such disturbances it must reflect a part of the energy of a diverging spherical wave. It is here shown that this property is possessed by a medium, such as that proposed by Kelvin, in which the elastic forces are of gyrostatic origin. This is due to the fact that, for a small constant angular displacement of an element of this medium, the restoring torque, instead of being constant, decreases progressively with time.

Introduction

IN A companion paper[1] it is pointed out that it may be possible to describe the behavior of material particles as that of moving patterns of wave motion, provided a medium can be found which is capable of sustaining a localized oscillatory disturbance. In most media this is not possible, for the energy of the disturbance would be propagated away in all directions. Something special in the way of a medium is therefore called for. It must be capable of trapping the wave energy released from a central source. Kelvin proposed a mechanical medium, the equations of which, for small disturbances, were identical with those of Maxwell for free space. The medium derived its elasticity from gyrostats. He recognized that, for finite disturbances, the restoring torque depends on the time as well as the angular displacement. It is the present purpose to show that this time dependence imparts to his medium exactly the energy trapping property required.

The Gyrostatic Ether

The concept of an ether with stiffness to rotation originated with MacCullagh[2] in 1839, and was further developed by Kelvin[3] in 1888. MacCullagh showed that certain optical phenomena associated with reflection could not be represented by the elastic solid ether of Fresnel, but required for their mechanical representation a medium in which the potential energy is a function of what is now called the curl of the displacement. Fitzgerald[4] remarked in 1880 that its equations are identical with those of the electromagnetic

[1] R. V. L. Hartley, Matter, a Mode of Motion—this issue of the *Bell System Technical Journal*.

[2] Collected Works of James MacCullagh, Longmans Green & Co., London, 1880, p. 145.

[3] Mathematical and Physical Papers of Sir William Thomson, Vol. III, Art. XCIX, p. 436, and Art. C, p. 466.

[4] Phil. Trans. 1880, quoted by Larmor, Ether and Matter, Cambridge Univ. Press, 1900, p. 78.

theory of optics developed by Maxwell. This conclusion is confirmed in later discussion by Gibbs,[5] Larmor,[4] and Heaviside.[6]

Kelvin, apparently unaware of MacCullagh's work, was led by similar considerations to the same result. He went farther and devised a physical model which consisted of a lattice, the points of which were connected by extensible, massless, rigid rods in such a manner that the structure as a whole was incompressible and non-rigid. Each of these rods supported a pair of oppositely rotating gyrostats. By a gyrostat he meant a spinning rotor mounted in a gimbal so that it is effectively supported at its center of mass and can have its spin axis rotated by a rotation of the mounting. The resultant angular momentum of the rotors was the same in all directions.

This model, considered as a continuous medium, exhibits a stiffness to absolute rotation, the nature of which can be described by comparing it with the elasticity of a solid. A solid is characterized by a rigidity n such that small displacements u, v, w are accompanied by a stress tensor, one component of which is

$$n\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right).$$

For the ether model the corresponding component is

$$n\left(\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right)=2n\varphi$$

where φ is a small angular displacement of the element about the z axis. More generally a small vector rotation $\overline{\Delta\varphi}$ is accompanied by a vector restoring torque per unit volume,

$$\overline{\Delta T}=-4n\overline{\Delta\varphi}. \tag{1}$$

The quantity $4n$ therefore represents a stiffness to angular displacement of the element.

In the appendix it is shown that the lattice of gyrostats, treated as a continuous medium, exhibits this kind of elasticity. It is also shown that for infinitesimal displacements, the medium is described by the wave equations (8a and 6a).

$$\nabla\times\left(\frac{\overline{T}}{2}\right)=\rho_0\frac{\partial\bar{q}}{\partial t}, \tag{2}$$

$$\nabla\times\bar{q}=-\frac{1}{\eta_0}\frac{\partial}{\partial t}\left(\frac{\overline{T}}{2}\right), \tag{3}$$

[5] Collected Works of J. Willard Gibbs, Longmans Green & Co., New York 1928, Vol. II, p. 232.

[6] Heaviside, Electromagnetic Theory, Ernest Benn, Ltd., London, 1893, Vol. I, p. 226.

where ρ_0 is the constant density, η_0 is a generalized stiffness of the undisturbed medium, given by (7a), $\bar{q}$ is the vector velocity, and $\overline{T}$ is the torque per unit volume. In a plane wave $\bar{q}$ is normal to the direction of propagation. $\frac{\overline{T}}{2}$ is a tractive force per unit area in the direction of $\bar{q}$, which acts on a surface normal to the direction of propagation.

If, however, the amplitude is finite the equations become much more complicated. For present purposes we need consider only waves for which there is no component of velocity or torque in the direction of propagation, and we need consider only plane polarized waves for which the direction of the velocity is the same at all times and places. Also, as will appear below, we are concerned with the equations which describe a wave of infinitesimal amplitude which is superposed on a finite disturbance. This description need cover only infinitesimal ranges of time and position. It can therefore be expressed in terms of wave equations in which the constants of the medium have local instantaneous values which depend on the finite disturbance.

Subject to these restrictions it is shown in the appendix that (2) is to be replaced by (23a)

$$\nabla \times \left(\frac{\overline{T}}{2}\right) = \bar{l}_q \rho \frac{\partial q}{\partial t}, \tag{4}$$

where $\bar{l}_q$ is a unit vector in the fixed direction of the velocity, and ρ is an instantaneous local density, defined in terms of the finite disturbance by (20a). And, in place of (3), (22a)

$$\nabla \times \bar{q} = -\bar{l}_\varphi \frac{1}{\rho c^2}\left(\frac{\partial}{\partial t}\left(\frac{\overline{T}}{2}\right) + 2\frac{\partial f}{\partial t}\right), \tag{5}$$

where $\bar{l}_\varphi$ is a unit vector in the direction of the axis of rotation, ρ is again an instantaneous local density, c is an instantaneous local velocity derived in the usual way from ρ and an instantaneous local stiffness η, while f is a function defined by the relation, (13a),

$$\overline{T} = -\bar{l}_\varphi 4 f(\varphi, t).$$

This function takes account of the fact that when the spin axis of the rotor is given a constant finite displacement, the restoring torque is not constant as in (1), but changes with time as the spin axis rotates toward the axis of displacement, and so reduces the component of the spin which is normal to the displacement axis and so is effective in producing stiffness. $-4\frac{\partial f}{\partial t}$ represents the rate of this change in torque for a fixed angular displacement. $-4\frac{\partial f}{\partial \varphi}$ is to be interpreted as the rate of change of torque with angular

displacement, when the time consumed is infinitesimal, that is when the angular velocity is infinite. It is therefore an instantaneous local angular stiffness from which the instantaneous local generalized stiffness η is derived as in (19a).

To simplify these expressions, let the direction of propagation be x and that of q be y. Then

$$\nabla \times \bar{q} = i \frac{\partial}{\partial x} (jq) = k \frac{\partial q}{\partial x},$$

so $\bar{l}_\varphi$ is in the direction of z, and represents a clockwise rotation about z. (5) then becomes the scaler equation

$$\frac{\partial q}{\partial x} = - \frac{1}{\rho c^2} \left[\frac{\partial}{\partial t} \left(\frac{T}{2} \right) + 2 \frac{\partial f}{\partial t} \right]. \tag{6}$$

T is also in the z direction, so

$$\nabla \times \left(\frac{T}{2} \right) = i \frac{\partial}{\partial x} \left(k \frac{T}{2} \right) = - j \frac{\partial}{\partial x} \left(\frac{T}{2} \right).$$

But $\bar{q}$ is in the y direction, so

$$\frac{\partial}{\partial x} \left(\frac{T}{2} \right) = - \rho \frac{\partial q}{\partial t}. \tag{7}$$

These, then, are the desired equations of motion, for the type of wave under consideration.

The Generation of Reflected Waves

In this section we shall show that when a finite wave is propagated in this medium each element of the medium becomes the source of auxiliary waves which propagate in both directions from the source.

To do this we shall make use of the argument by which Riemann[7] showed that this does not occur for sound waves in an ideal gas. This will first be restated in more modern language. We consider a plane wave propagating along the x axis. We picture the finite pressure p and the longitudinal velocity u at a point in the medium as having been built up by the successive superposition of waves of infinitesimal amplitude, each propagating relative to the medium in its condition at the time of its superposition. If the first increment is propagating in the positive direction,

$$du = \frac{dp}{\rho c},$$

[7] Lamb, Hydrodynamics, Sixth Edition, p. 481. Rayleigh, Theory of Sound, Second Edition, Vol. II, p. 38.

where the characteristic resistance is ρc. Here

$$c^2 = \frac{dp}{d\rho}.$$

He assumes adiabatic expansion, so that p and c are functions of ρ only. If a second incremental wave of pressure dp, also traveling in the positive direction, be added, its velocity increment, being relative to the medium, will add to that already present. Its value will be related to dp through a new characteristic resistance corresponding to the modified density resulting from the previous increment. Hence the velocity u resulting from a large number of such waves will be

$$u = \int_0^p \frac{dp}{\rho c} = w,$$

where w is the quantity represented by ω in Lamb's version. If, then, all of the wave propagation is in the positive direction

$$u = w.$$

Similarly, if an incremental wave is traveling in the negative direction,

$$du = \frac{-dp}{\rho c},$$

and the condition for all the propagation to be in that direction is

$$u = -w.$$

Obviously, then, if u has some other value than one of these it results from the addition of increments some of which propagate in each direction.

Riemann deduces from the aerodynamic equations that

$$\left(\frac{\partial}{\partial t} + (u + c)\frac{\partial}{\partial x}\right)(w + u) = 0, \tag{8}$$

$$\left(\frac{\partial}{\partial t} + (u - c)\frac{\partial}{\partial x}\right)(w - u) = 0, \tag{9}$$

That is, the value of $w + u$ is propagated in the positive direction with a velocity of $c + u$ and that of $w - u$, in the negative direction with a velocity $c - u$. If, over a finite range of x, a disturbance be set up such that neither of these quantities is zero, it must be made up of incremental waves in both directions. However, as $w + u$ propagates positively it will be accompanied at any instant by a value of $w - u$ which has been propagated from the other direction. But, since the value of this was initially finite over a limited distance only, when all of this finite range is passed, $w - u$ will be zero, u will

be equal to w and all of the wave will be traveling positively. A similar argument applies at the negative side of the wave. Thus the initial disturbance breaks up into two parts which travel in opposite directions without reflection. More generally, these considerations hold for any medium in which the stress is a function of the strain only.

For the ether model, since we have assumed the displacements are normal to the direction of propagation, the velocity of wave propagation relative to the medium is the same as that relative to the axes.

If now, following Riemann, we let

$$dw = \frac{1}{\rho c}\, d\left(\frac{T}{2}\right), \tag{10}$$

so that now

$$w = \int \frac{1}{\rho c}\, d\left(\frac{T}{2}\right),$$

then from (7) and (6)

$$\frac{\partial q}{\partial t} = -\, c\, \frac{\partial w}{\partial x},$$

$$\frac{\partial w}{\partial t} = -\, c\, \frac{\partial q}{dx} - \frac{2}{\rho c}\, \frac{\partial f}{dt}.$$

Adding and subtracting gives

$$\left(\frac{\partial}{\partial t} + c\, \frac{\partial}{\partial x}\right)(w + q) = -\, \frac{2}{\rho c}\, \frac{\partial f}{\partial t},$$

$$\left(\frac{\partial}{\partial t} - c\, \frac{\partial}{\partial x}\right)(w - q) = -\, \frac{2}{\rho c}\, \frac{\partial f}{\partial t},$$

which are to be compared with (8) and (9). Hence when $\frac{\partial f}{\partial t}$ is not zero the values of $w + q$ and $w - q$ are not propagated without change.

To show that reflection occurs, consider a disturbance at a point x at time t, characterized by q and w. At x and $t + \Delta t$, $w + q$ will differ from the value it had at $x - c\Delta t$, t, or $w + q - \frac{\partial}{\partial x}(w + q)c\Delta t$, by $-\frac{2}{\rho c}\frac{\partial f}{\partial t}\Delta t$. The increment at x in time Δt is

$$\Delta w + \Delta q = -\, \frac{\partial}{\partial x}(w + q)c\Delta t - \frac{2}{\rho c}\, \frac{\partial f}{\partial t}\, \Delta t,$$

and

$$\Delta w - \Delta q = \frac{\partial}{\partial x}(w - q)c\Delta t - \frac{2}{\rho c}\frac{\partial f}{\partial t}\Delta t.$$

From which

$$\Delta w = -c\frac{\partial q}{\partial x}\Delta t - \frac{2}{\rho c}\frac{\partial f}{\partial t}\Delta t,$$

$$\Delta q = -c\frac{\partial w}{\partial x}\Delta t.$$

Hence the velocity is the same as when $\frac{\partial f}{\partial t}$ is zero but w is changed by $-\frac{2}{\rho c}\frac{\partial f}{\partial t}\Delta t$. But the only way in which w can change with q constant is by adding waves of equal amplitude propagating in opposite directions, so that their contributions to w are equal and those to q are equal and opposite. From (10) this involves an increment of $\frac{T}{2}$ of $-2\frac{\partial f}{\partial t}\Delta t$ or a time rate of change of $-2\frac{\partial f}{\partial t}$. This agrees with (6), from which it is evident that the presence of $\frac{\partial f}{\partial t}$ alters $\frac{\partial q}{\partial x}$ from what it would otherwise be by $-\frac{2}{\rho c^2}\frac{\partial f}{\partial t}$. But, since q is unchanged, the velocities at $x + \frac{\Delta x}{2}$ and $x - \frac{\Delta x}{2}$ are increased by $-\frac{1}{\rho c^2}\frac{\partial f}{\partial t}\Delta x$ and $\frac{1}{\rho c^2}\frac{\partial f}{\partial t}\Delta x$. The first is the velocity associated with an auxiliary wave which propagates in the positive direction of x, and the second that of one which propagates in the negative direction, that is a reflected wave. Hence the medium generates a reflected wave of $\frac{1}{\rho c^2}\frac{\partial f}{\partial t}$ per unit length in the direction of propagation.

The Reflection of a Progressive Diverging Wave

So far attention has been confined to a single point. If a continuous disturbance is being propagated, it is important to know how the waves reflected at different points combine, for it is conceivable that they may interfere destructively. From the standpoint of the application to be made of these results in a companion paper, the case of most interest is that in which energy is propagated outward from a central generator as a sinusoidal wave of finite amplitude, beginning at time zero. Near the center, the wave of displacement will include radial as well as tangential components. As the radius

increases the radial components become relatively negligible. We shall confine our attention to this outer region, where, in the absence of reflection, the propagation differs from that of a plane wave only in that the amplitude varies inversely as the radius. We shall neglect the effect of any reflections on the outgoing wave, and calculate the resultant reflected wave at a radius r_1 as a function of the time and so of the radial distance r the wave front has traveled.

If the outgoing wave were of infinitesimal amplitude, its velocity q_0 could be represented by

$$q_0 = \frac{r_0}{r} Q_0 \sin(\omega t - kr), \tag{11}$$

for values of $r < ct$, and by zero for $r > ct$, where Q_0 is the amplitude at some reference radius r_0. The sine function is chosen to avoid the necessity of an infinite acceleration at the wave front, as would be required by a cosine function. When the amplitude is finite this wave suffers distortion due to the fact that k which is equal to $\frac{\omega}{c}$ varies slightly with the variations in the instantaneous value of c. However, these will be small and, since fluctuations in velocity alone do not cause reflection, we shall neglect them. The procedure is to make use of q_0 to calculate the reflected wave increment generated in a length $\Delta r'$ at a radius r', calculate the amplitude and phase of this at a fixed point $r_1 < r'$, and at r_1 integrate the waves received there for values of r' from r_1 to the farthest point from which reflected waves can reach r_1 at the time t under consideration.

To find the reflected wave generated in a length $\Delta r'$ at r', we have from above that its velocity

$$\Delta q' = \frac{1}{\rho c^2} \frac{\partial f}{\partial t} \Delta r'.$$

From (21a), (19a) and (17a)

$$\frac{1}{\rho c^2} = \frac{F_1'}{\eta_0 \left(1 - a\left[\int \varphi \, dt\right]^2\right)},$$

where η_0 and a are constants of the medium given by (7a) and (15a). From (18a)

$$\frac{\partial f}{\partial t} = -a\eta_0 \varphi^2 \int \varphi \, dt,$$

so

$$\frac{dq'}{dr'} = -\frac{aF_1'\varphi^2 \int \varphi\, dt}{1 - a\left[\int \varphi\, dt\right]^2}$$

which reduces to

$$\frac{dq'}{dr'} = -a\varphi^2 \int \varphi\, dt,$$

if we neglect second powers of the variables compared with unity.

To the same accuracy, from (14a)

$$\varphi = \frac{1}{2} \int \frac{\partial q_0}{\partial r'}\, dt.$$

From (11)

$$\frac{\partial q_0}{\partial r'} = -\frac{r_0 Q_0}{r'}\left[k \cos(\omega t - kr') + \frac{1}{r'} \sin(\omega t - kr')\right].$$

Here k is 2π over the wavelength so, if as we have assumed r_1, and therefore also r', is large compared with the wavelength, we may neglect the second term. Then

$$\varphi = -\frac{r_0 Q_0}{2cr'} \sin(\omega t - kr'),$$

$$\int \varphi\, dt = \frac{r_0 Q_0}{2c\omega r'} \cos(\omega t - kr'),$$

$$\frac{dq'}{dr'} = -\frac{a}{8\omega}\left(\frac{r_0 Q_0}{cr'}\right)^3 \sin^2(\omega t - kr') \cos(\omega t - kr'),$$

$$= -\frac{a}{8\omega}\left(\frac{r_0 Q_0}{cr'}\right)^3 [\cos(\omega t - kr') + \cos 3(\omega t - kr')].$$

This, when multiplied by $\Delta r'$, gives the value at r' of the wave, generated in the interval $\Delta r'$, which propagates in the negative direction of r. This is made up of components of frequency ω and 3ω. We are primarily interested, from the stand-point of reflection, in that of frequency ω, so we shall confine our attention to this component, with the understanding that the other can be treated in exactly the same fashion. As the fundamental component propagates inward to r_1 it increases in amplitude in the ratio $\frac{r'}{r_1}$ and suffers a phase lag of $k(r' - r_1)$. If we call the resultant of all the reflected waves at r_1, q_1', then the contribution to q_1' of the wave generated at r' is

$$\Delta q_1' = -\frac{a}{8\omega}\left(\frac{r_0 Q_0}{c}\right)^3 \frac{1}{r_1 r'^2}\cos(\omega t + kr_1 - 2kr')\Delta r'.$$

This is to be integrated from r_1 to the farthest point from which a reflected wave has reached r_1 at the instant t under consideration. This point is at $\frac{1}{2}(r_1 + ct)$. So

$$q_1' = -\frac{a}{8r_1\omega}\left(\frac{r_0 Q_0}{c}\right)^3 \int_{r_1}^{\frac{1}{2}(r_1+ct)} \frac{1}{r'^2}\cos(\omega t + kr_1 - 2kr')\,dr'.$$

Here the integrand is a function of r' and t and the upper limit of integration is also a function of t. We therefore make use of the relation[8]

$$\frac{d}{d\alpha}\int_a^b f(x,\alpha)\,dx = \int_a^b \left(\frac{\partial}{\partial\alpha} f(x,\alpha)\right) dx + f(b,\alpha)\frac{db}{d\alpha} - f(a,\alpha)\frac{da}{d\alpha}.$$

Putting t for α, r' for x we have

$$\frac{dq_1'}{dt} = \frac{a}{8r_1}\left(\frac{r_0 Q_0}{c}\right)^3 \left[\int_{r_1}^{\frac{1}{2}(r_1+ct)} \frac{1}{r'^2}\sin(\omega t + kr_1 - 2kr') - \frac{2c}{\omega}\frac{1}{(r_1+ct)^2}\right]$$

which, upon integration becomes,

$$\frac{dq_1'}{dt} = \frac{a}{8r_1}\left(\frac{r_0 Q_0}{c}\right)^3 \left(\frac{1}{r_1} si(\omega t + kr_1) - 2k[Si(\omega t - kr_1) - Si(2kr_1)] \cdot \sin(\omega t + kr_1) - [Ci(\omega t + kr_1) - Ci(2kr_1)] \cdot \cos(\omega t + kr_1) - \frac{2c}{\omega(r_1+ct)^2}\right).$$

Since q_1' is zero when t is $\frac{r_1}{c}$, its value at t will be found by integrating from $\frac{r_1}{c}$ to t, so

$$q_1' = \frac{a}{8r_1^2\omega}\left(\frac{r_0 Q_0}{c}\right)^3 \left(-\cos(\omega t - kr_1) + \frac{2r_1}{r_1+ct} - 2kr_1 \cdot \left[\omega\int_{r_1/c}^{t} Si(\omega t + kr_1)\sin(\omega t + kr_1)\,dt + Si(2kr_1) \cdot [\cos(\omega t + kr_1) - \cos 2kr_1] - \omega\int_{r_1/c}^{t} Ci(\omega t + kr_1)\cos(\omega t + kr_1)\,dt + Ci(2kr_1)[\sin(\omega t + kr_1) - \sin 2kr_1]\right]\right).$$

which reduces to

[8] Byerly, Integral Calculus, second edition p. 99.

$$q_1' = -\frac{a}{8r_1^2\omega}\left(\frac{r_0 Q_0}{c}\right)^3\left(\cos(\omega t - kr_1) - \frac{2r_1}{r_1 + ct} + 2kr_1\right.$$
$$\cdot\,[-[Si(\omega t + kr_1) - Si(2kr_1)]\cos(\omega t + kr_1)$$
$$- [Ci(\omega t + kr_1) - Ci(2kr_1)]\sin(\omega t + kr_1)$$
$$\left. + Si(2\omega t + 2kr_1) - Si(4kr_1)]\right).$$

The first term represents the value at r_1 of an outwardly moving wave in phase quadrature with the main wave. The second is a transient, the value of which is equal and opposite to that of the first term at the instant that the main wave passes r_1. The first two terms in the inner bracket are waves which propagate inward and so are to be regarded as reflections of the main wave. The last two terms represent a velocity which is zero when the main wave passes r_1, and subsequently oscillates about and approaches $\frac{\pi}{2} - \mathrm{Si}(4kr_1)$. Physically it appears to result from the particular form chosen for the main wave, which starts abruptly as a sine wave. The time integral of the impressed force, and so the applied momentum, has a component in one direction. Presumably if the main wave built up gradually these terms would be absent.

Returning to the reflected waves, their amplitudes are zero when the main wave passes r_1, after which they become finite. $Si(x)$ and $Ci(x)$ oscillate about and approach $\frac{\pi}{2}$ and zero respectively as x approaches infinity. Hence, as t increases indefinitely, the amplitudes of the reflected waves approach $\frac{\pi}{2} - Si(2kr_1)$ and $Ci(2kr_1)$. For the assumed large values of $2kr_1$ these quantities are small compared with unity. When multiplied by $2kr_1$ their variation is very slow. Hence the amplitudes vary roughly as $\frac{1}{r_1^2}$, and approach zero as the main wave at r_1 approaches an ideal plane one.

However, the significant fact is not that the reflected waves are small but that they are of finite magnitude. Because of this the main wave will not behave exactly as we assumed above, but will decrease slightly more rapidly with increasing radius. This should increase the reflection slightly, for the existence of the reflected wave is dependent on the decrease in amplitude with distance when the radius of curvature is finite.

To describe exactly what happens when the generator begins sending out waves from a central point would be hopelessly complicated, but we may form a general picture. In the early stages where the curvature is considerable, the reflected waves would be quite large and the main wave would be

correspondingly attenuated. The arrival of the reflected waves at the generator adds a reactive component to the impedance of the medium, as seen from the generator, which reduces the power delivered to the medium. Meanwhile energy is being stored as standing waves in the medium and the rate of flow of energy in the wavefront is decreasing. The energy in successive shells of equal radial thickness decreases with increasing r, instead of being uniform as it would be in the absence of reflection. In the limit it approaches zero, but as the rate of decrease depends on the curvature, the rate of approach also approaches zero. As the rate at which energy is stored and that at which it is carried outward at the wavefront both approach zero, the resistance which the medium offers to the generator approaches zero, and its impedance approaches a pure reactance.

The total energy stored in the medium depends on how the over-all attenuation of the main wave is related to its amplitude. If there were no attenuation, the impedance would remain a pure resistance, the energy in successive shells would all be the same, and the total energy would increase linearly with r, and so with the time, and approach infinity. If the attenuation were independent of r, the total energy would approach a finite value. The present case is intermediate between these, the attenuation being finite but approaching zero with increasing r. If we assume it to vary as some power of the amplitude of the velocity, then W. R. Bennett has shown that if this power is less than the first the total energy approaches a finite value. If it is equal to the first, the energy approaches infinity as $\log r$, and if it is greater than this, the power approaches infinity more rapidly. Until more is known as to the actual variation of amplitude with distance, nothing definite can be said about the limit of the total energy.

APPENDIX: EQUATIONS OF THE KELVIN ETHER

We are concerned with the wave properties of the model for wavelengths long enough compared with the lattice constant so that it may be regarded as a continuous medium. Its density is equal to the average mass of the gyrostats per unit volume. Its elastic properties are to be derived from the resultant of the responses of the individual gyrostats.

We shall therefore begin by considering the behavior of a single element, which is shown schematically in Fig. 1. Here the outer ring of the gimbal, which is rigidly connected with the lattice, lies in the $x\,y$ plane. The axis about which the inner ring rotates is in the x direction, and the spin axis C of the rotor is in the z direction. We wish to examine the effect of a small angular displacement φ of the lattice, that is, of the outer ring. If it is about x or z, it will, because of the frictionless bearings, make no change in the rotor. If it is about y it will produce an equal displacement of the spin axis

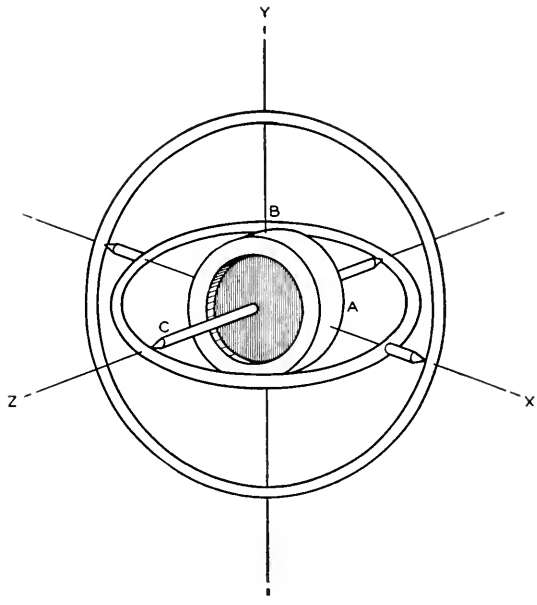

Fig. 1—Diagram of a gyrostat, showing its axes of rotation.

C about y. To study its effect we make use of Euler's equations for a rotating rigid body.[9]

$$A\frac{d\omega_1}{dt} - (B - C)\omega_2\omega_3 = L,$$

$$B\frac{d\omega_2}{dt} - (C - A)\omega_3\omega_1 = M,$$

$$C\frac{d\omega_3}{dt} - (A - B)\omega_1\omega_2 = N,$$

where ω_1, ω_2 and ω_3 are the angular velocities about three principal axes of inertia, fixed in the rotor, the moments of inertia about which are A, B and C, and L, M, and N are the accompanying torques about the three axes. They are also at any instant the values of the torques about that set of axes, fixed in space, which, at the instant, coincide with the axes 1, 2, 3, which are fixed relative to the body. We let the 3 axis coincide with the spin axis C. We choose as the 1 and 2 axes, lines in the rotor which, at the instant, are in the x and y directions respectively. Since the moments of

[9] Jeans, Theoretical Mechanics, Ginn and Co., p. 308.

inertia about these are equal, A and B are equal. By virtue of the frictionless bearings the external torques L and N about 1 and 3 are zero.

Introducing these relations we have

$$A\frac{d\omega_1}{dt} + (C - A)\omega_2\omega_3 = 0, \tag{1a}$$

$$A\frac{d\omega_2}{dt} - (C - A)\omega_1\omega_3 = M, \tag{2a}$$

$$C\frac{d\omega_3}{dt} = 0. \tag{3a}$$

From (3a) the velocity of spin ω_3 remains constant. The torque M about y is then to be found from (1a) and (2a). For very small displacements,

$$\omega_2 = \dot{\varphi}.$$

Putting this in (1a) and integrating from zero to t, assuming φ to be zero at $t = 0$, gives

$$\omega_1 = -\frac{C - A}{A}\omega_3\varphi.$$

(2a) then becomes

$$A\ddot{\varphi} + \frac{(C - A)^2}{A}\omega_3^2\varphi = M.$$

This represents an angular inertia A and stiffness $\frac{(C - A)^2\omega_3^2}{A}$. The system will therefore resonate at a frequency $\frac{(C - A)\omega_3}{A}$. If the frequencies involved in the variation of φ are small compared with this, the inertia torque will be negligible, and the system will behave as a stiffness. If the displacements about A associated with ω_1 are very small the restoring torque M will act substantially about the y axis. That is, the lattice will encounter a stiffness to rotation.

Since the large number of gyrostats in an element of the model are oriented in all directions, an angular displacement of the lattice about y will generally not be about the B axis for each gyrostat. If it makes an angle α with this axis, then only the component $\varphi\cos\alpha$ of the angular displacement will be transmitted to the rotor. The resulting torque will then be $S\cos\alpha$, where

$$S = \frac{(C - A)^2\omega_3^2}{A}.$$

It will be directed about B and so will not be parallel to the applied displacement. However, if a second gyrostat has the position which the first

would have if it were rotated about y through π, its torque along y is the same as that of the first, and that normal to it is equal and opposite. Hence, if the gyrostats are properly oriented, the resultant torque will be parallel to the displacement and the medium will be isotropic. The y component of the opposing torque will be $S\varphi \cos^2 \alpha$. Thus if the B axes are uniformly distributed in space the total torque will be one third what it would be if they were all parallel to the axis of the applied displacement. Hence if there are N gyrostats per unit volume the vector restoring torque $\bar{T}$ per unit volume will be

$$\bar{T} = -\frac{N}{3} \frac{(C - A)^2 \omega_3^2}{A} \bar{\varphi}. \tag{4a}$$

The next step is to derive the wave equations for a medium having this stiffness to rotation. If the vector velocity $\bar{q}$ is very small,

$$\nabla \times \bar{q} = 2 \frac{\partial \bar{\varphi}}{\partial t}, \tag{5a}$$

where $\bar{\varphi}$ is a vector angular displacement of an element of the medium at the point under consideration. 2φ plays a role analagous with that of the dilatation in compressional waves. Then, from (4a) and (5a),

$$\nabla \times \bar{q} = -\frac{1}{\eta_0} \frac{\partial}{\partial t} \left(\frac{\bar{T}}{2} \right), \tag{6a}$$

where the generalized stiffness of the undisturbed medium,

$$\eta_0 = \frac{N}{12} \frac{(C - A)^2}{A} \omega_3^2. \tag{7a}$$

To get the companion equation, we interpret the torque exerted by an element in terms of the forces it exerts on the surfaces of neighboring elements. Let the x axis Fig. 2 be in the direction of the torque $T\Delta x^3$ which is exerted by the medium within the small cube. This very small torque can be resolved into the sum of two couples, one consisting of an upward force $F_y \Delta x^2$ on the right face and an equal downward force on the left one, and the other of a leftward force $F_z \Delta x^2$ on the upper surface and a rightward one on the lower one. But, if there is not to be a shearing stress, F_y and F_z must be equal, and each equal to $\frac{T}{2}$. Thus a torque per unit volume T is equivalent to a set of tangential surface forces per unit area of $\frac{T}{2}$ each.

Now consider the force exerted on an element by its neighbors, through the adjoining surfaces. To take the simplest case, let T in Fig. 2 be everywhere in the x direction and independent of z but varying with y. Then

the forces exerted on the upper and lower surfaces are equal and opposite. That downward on the right face exceeds that upward on the left by $\frac{\partial}{\partial y}\left(\frac{T}{2}\,\Delta x^2\right)\Delta y$, so the force in the z direction is $-\frac{\partial}{\partial y}\left(\frac{T}{2}\right)\Delta x^3$. By extending the argument to three dimensions it is easily shown that the total

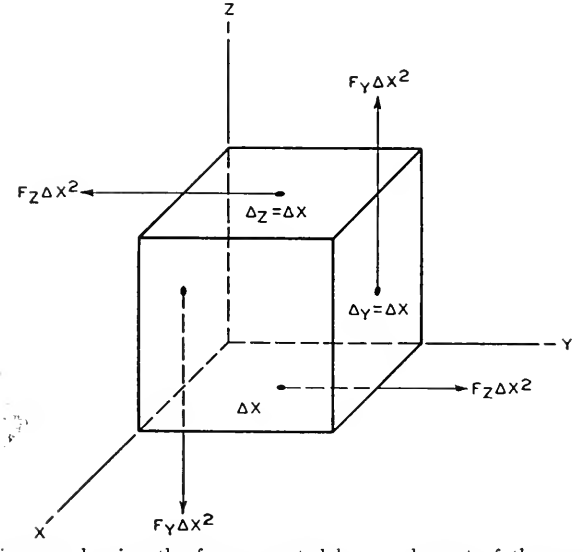

Fig. 2—Diagram showing the forces exerted by an element of the medium through its surfaces.

force is $\nabla \times \left(\frac{\overline{T}}{2}\right)\Delta x^3$. If ρ_0 is the density of the medium this force must equal $\rho_0 \Delta x^3 \frac{d\bar{q}}{dt}$, so

$$\nabla \times \left(\frac{\overline{T}}{2}\right) = \rho_0 \frac{d\bar{q}}{dt},$$

which, since $\bar{q}$ is small, reduces to

$$\nabla \times \left(\frac{\overline{T}}{2}\right) = \rho_0 \frac{\partial \bar{q}}{\partial t}. \tag{8a}$$

From this and (6a) the velocity of propagation is $(\eta_0/\rho_0)^{1/2}$ and the characteristic resistance is $(\rho_0\eta_0)^{1/2}$. In a plane wave the displacement is normal to the direction of propagation. The stress is a tractive force per unit area $\frac{T}{2}$ acting in a surface normal to the direction of propagation. It is in the direction of the velocity and in phase with it.

However, we are also interested in the case where the amplitudes are not negligible. We shall confine our attention to those cases where, as in plane or spherical waves at a distance from the source, the velocity is normal to the direction of propagation and the variations in the plane of the wave front are negligible. (5a) then becomes much more complicated. $\nabla \times \bar{q}$ is, however, still a function of $\frac{\partial \bar{\varphi}}{\partial t}$, say $2F_1\left(\frac{\partial \bar{\varphi}}{\partial t}\right)$. Then, for small variations of $\frac{\partial \bar{\varphi}}{\partial t}$ in the neighborhood of a particular value, we may write

$$\nabla \times \bar{q} = 2F_1'\left(\frac{\partial \bar{\varphi}}{\partial t}\right)\frac{\partial \bar{\varphi}}{\partial t} \tag{9a}$$

where $F_1'\left(\frac{\partial \bar{\varphi}}{\partial t}\right)$ is a function of the particular value of $\frac{\partial \bar{\varphi}}{\partial t}$. This relation is to take the place of (5a). Similarly, if

$$\nabla \times \left(\frac{\overline{T}}{2}\right) = F_2\left(\frac{\partial \bar{q}}{\partial t}\right),$$

then, in place of (8a), we are to use, for small variations,

$$\nabla \times \left(\frac{\overline{T}}{2}\right) = F_2'\left(\frac{\partial \bar{q}}{\partial t}\right)\frac{\partial \bar{q}}{\partial t}. \tag{10a}$$

When we come to the transition from (5a) to (6a), however, the situation is somewhat different. To see how this comes about, we go back to the behavior of the single gyrostat of Fig. 1. It was assumed above that the B axis coincided with the y axis However, when the displacement of the rotor about A is finite, this is no longer exactly true. The situation is then as shown in Fig. 3. A rotation φ of the lattice about y displaces A in the $x\,z$ plane by φ. The accompanying rotation of the rotor about A causes B to make an angle θ with y, which is independent of φ. Then

$$\omega_2 = \frac{d\varphi}{dt}\cos\theta.$$

From (1a)

$$\omega_1 = -\frac{C - A}{A}\,\omega_3 \int \frac{d\varphi}{dt}\cos\theta\, dt.$$

Also

$$\theta = \int \omega_1\, dt,$$

$$= -\frac{C - A}{A}\,\omega_3 \iint \frac{d\varphi}{dt}\cos\theta\, dt\, dt, \tag{11a}$$

which determines θ as a function of φ and t. From (2a), neglecting the first term as above,

$$M = S \int \frac{d\varphi}{dt} \cos \theta \, dt,$$

and the restoring torque about y, or

$$T_y = -S \cos \theta \int \frac{d\varphi}{dt} \cos \theta \, dt. \tag{12a}$$

This, together with (11a), determines T_y as a function of φ and t, instead of φ alone as it is for infinitesimal displacements.

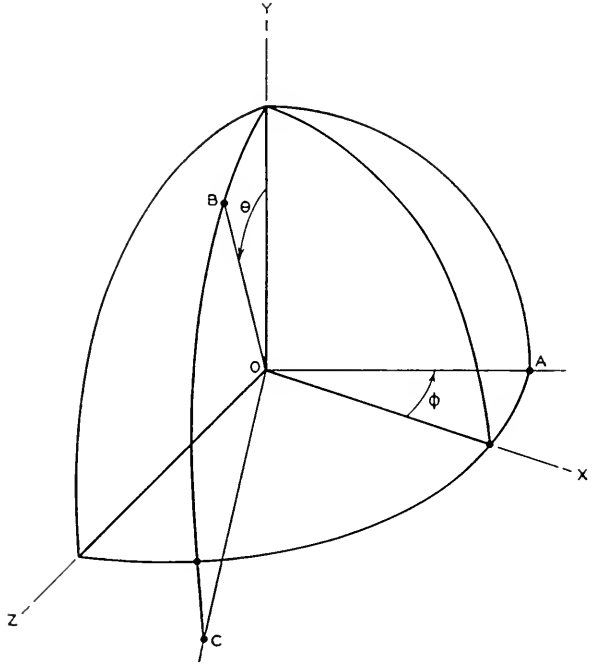

Fig. 3—Diagram showing the displacement of the axes of a gyrostat.

We assumed here that, in the rest position of the rotor, its B axis coincides with that of the applied displacement φ. When this is not the case, the relations are more complicated, but they should be qualitatively the same. Hence, for an element of the medium, the torque per unit volume should be a function of φ and t similar to T_y, which reduces to $-4\eta_0\varphi$ for very small displacements. Since the restoring torque is in the direction of φ we may

write

$$\overline{T} = -\bar{l}_\varphi 4f(\varphi, t) \tag{13a}$$

where $\bar{l}_\varphi$ is a unit vector in the direction of the axis of rotation.

The derivation of the wave equation is much simpler if we consider only the case of present interest where the direction of the rotation is everywhere the same so that $\bar{l}_\varphi$ is constant. Then (9a) can be written as

$$\nabla \times \bar{q} = \bar{l}_\varphi 2F_1' \left(\frac{\partial \varphi}{\partial t}\right) \frac{\partial \varphi}{\partial t}, \tag{14a}$$

and (13a) as

$$T = -4f(\varphi, t).$$

We wish now to replace $\dfrac{\partial \varphi}{\partial t}$ by $\dfrac{\partial}{\partial t}\left(\dfrac{T}{2}\right)$. These partial derivatives refer to a constant position so we are interested in the total time derivatives of T as given by (12a). To get the desired relation we need to express T explicitly in terms of φ and t, that is, we must evaluate φ. Since the variables are small, we neglect their products of higher order than the third. Then

$$\cos\theta = 1 - \frac{1}{2} a \left[\int \varphi \, dt\right]^2,$$

where

$$a = \left(\frac{C - A}{A} \omega_3\right)^2. \tag{15a}$$

Putting

$$T = -4\eta_0 \cos\theta \int \frac{d\varphi}{dt} \cos\theta \, dt,$$

in accordance with (12a) and substituting for $\cos\theta$ gives

$$T = -4\eta_0 \left[\varphi - a\varphi \left[\int \varphi \, dt\right]^2 + a \int \varphi^2 \left(\int \varphi \, dt\right) dt\right].$$

Then

$$\frac{dT}{dt} = -4\eta_0 \left[\left(1 - a \left[\int \varphi \, dt\right]^2\right) \frac{d\varphi}{dt} - a\varphi^2 \int \varphi \, dt\right].$$

When φ is constant the first term is zero, so the second term can be interpreted as the partial derivative of T with respect to t. Physically this describes the change in torque for a fixed displacement which results from the

fact that, as the axis of the rotor rotates toward that of the applied torque, the component of the spin which is normal to the axis of displacement progressively diminishes. To interpret the first term, we let $\frac{d\varphi}{dt}$ increase indefinitely. The second term then becomes negligible, and when we divide through by $\frac{d\varphi}{dt}$, the left side becomes $\frac{dT}{d\phi}$. But the time increment which accompanies a finite increment of φ is now infinitesimal, and so this may be called the partial with respect to φ, with t constant.

We have then

$$\frac{dT}{dt} = -4\left(\frac{\partial f}{\partial \varphi}\frac{d\varphi}{dt} + \frac{\partial f}{\partial t}\right) \tag{16a}$$

where

$$\frac{\partial f}{\partial \varphi} = \eta_0\left(1 - a\left[\int \varphi\, dt\right]^2\right), \tag{17a}$$

$$\frac{\partial f}{\partial t} = -a\eta_0\varphi^2 \int \varphi\, dt. \tag{18a}$$

Substituting for $\frac{\partial \varphi}{\partial t}$ from (16a) in (14a),

$$\nabla \times \bar{q} = -\bar{l}_\varphi \frac{F_1'}{\frac{\partial f}{\partial \varphi}}\left(\frac{\partial}{\partial t}\left(\frac{T}{2}\right) + 2\frac{\partial f}{\partial t}\right).$$

We may interpret $\frac{\partial f}{\partial \varphi}$ as an instantaneous stiffness to rotation and define an instantaneous local generalized stiffness by the relation

$$\eta = \frac{\frac{\partial f}{\partial \varphi}}{F_1'}. \tag{19a}$$

Similarly from (10a) we may define an instantaneous density by the relation

$$\rho = F_2'. \tag{20a}$$

Then we may speak of an instantaneous velocity c given by

$$c^2 = \frac{\eta}{\rho}, \tag{21a}$$

and an instantaneous characteristic resistance ρc. Then

$$\nabla \times \bar{q} = -\bar{l}_\varphi \frac{1}{\rho c^2} \left(\frac{\partial}{\partial t} \left(\frac{T}{2} \right) + 2 \frac{\partial f}{\partial t} \right). \tag{22a}$$

(10a) becomes

$$\nabla \times \left(\frac{T}{2} \right) = \bar{l}_q \rho \frac{\partial q}{\partial t}, \tag{23a}$$

where $\bar{l}_q$ is a unit vector in the fixed direction of the velocity. These are the equations of motion which apply to a very small disturbance superposed on a finite disturbance.

Traveling-Wave Tubes

By J. R. PIERCE

[THIRD INSTALLMENT]

CHAPTER VII

EQUATIONS FOR TRAVELING-WAVE TUBE

Synopsis of Chapter

IN CHAPTER VI we have expressed the properties of a circuit in terms of its normal modes of propagation rather than its physical dimensions. In this chapter we shall use this representation in justifying the circuit equation of Chapter II and in adding to it a term to take into account the local fields produced by a-c space charge. Then, a combined circuit and ballistical equation will be obtained, which will be used in the following chapters in deducing various properties of traveling-wave tubes.

In doing this, the first thing to observe is that when the propagation constant Γ of the impressed current is near the propagation constant Γ_1 of a particular active mode, the excitation of that mode is great and the excitation varies rapidly as Γ is changed, while, for passive modes or for active modes for which Γ is not near to the propagation constant Γ_n, the excitation varies more slowly as Γ is changed. It will be assumed that Γ is nearly equal to the propagation constant Γ_1 of one active mode, is not near to the propagation constant of any other mode and varies over a small fractional range only. Then the sum of terms due to all other modes will be regarded as a constant over the range of Γ considered. It will also be assumed that the phase velocities corresponding to Γ and Γ_1 are small compared with the speed of light. Thus, (6.47) and (6.47a) are replaced by (7.1), where the first term represents the excitation of the Γ_1 mode and the second term represents the excitation of passive and "non-synchronous" modes. In another sense, this second term gives the field produced by the electrons in the absence of a wave propagating on the circuit, or, the field due to the "space charge" of the bunched electron stream. Equation (7.1) is the equation for the distributed circuit of Fig. 7.1. This is like the circuit of Fig. 2.3 save for the addition of the capacitances C_1 between the transmission circuit and the electron beam. We see that, because of the presence of these capacitances, the charge of a bunched electron beam will produce a field in addition to the field of a wave traveling down the circuit. This circuit is intuitively so appealing that it was originally thought of by guess and justified later.

Equation (7.1), or rather its alternative form, (7.7), which gives the voltage in terms of the impressed charge density, can be combined with the

ballistical equation (2.22), which gives the charge density in terms of the voltage, to give (7.9), which is an equation for the propagation constant. The attenuation, the difference between the electron velocity and the phase velocity of the wave on the circuit in the absence of electrons and the difference between the propagation constant and that for a wave traveling with the electron speed are specified by means of the gain parameter C and the parameters d, b and δ. It is then assumed that d, b and δ are around unity or smaller and that C is much smaller than unity. This makes it possible to neglect certain terms without serious error, and one obtains an equation (7.13) for δ.

In connection with (7.7) and Fig. 7.1, it is important to distinguish between the *circuit voltage* V_c, corresponding to the first term of (7.7), and the total voltage V acting on the electrons. These quantities are related by (7.14). The a-c velocity v and the convection current i are given within the approximation made ($C \ll 1$) by (7.15) and (7.16).

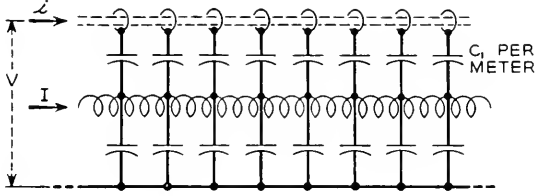

Fig. 7.1

7.1 Approximate Circuit Equation

From (6.47) we can write for a current $J = i$ and a summation over n modes

$$E_z = (1\ 2)(\Gamma^2 + \beta_0^2) i \sum_n \frac{(E^2/\beta^2 P)_n \Gamma_n^3}{(\Gamma_n^2 + \beta_0^2)(\Gamma_n^2 - \Gamma^2)} \tag{6.47a}$$

This has a number of poles at $\Gamma = \Gamma_n$. We shall be interested in cases in which Γ is very near to a particular one of these, which we shall call Γ_1. Thus the term in the expansion involving Γ_1 will change rapidly with small variations in Γ. Moreover, even if $(E^2/\beta^2 P)_1$ and Γ_1 have very small real components, $\Gamma_1^2 - \Gamma^2$ can be almost or completely real for values of Γ which have only small real components. Thus, one term of the expansion, that involving Γ_1, can go through a wide range of phase angles and magnitudes for very small fractional variations in Γ, fractional variations, as it turns out, which are of the order of C over the range of interest.

The other modes are either passive modes, for which even in a lossy circuit $(E^2/\beta^2 P)_n$ is almost purely imaginary, and Γ_n almost purely real,

or they are active modes which are considerably out of synchronism with the electron velocity. Unless one of these other active modes has a propagation constant Γ_n such that $|(\Gamma_1 - \Gamma_2)/\Gamma_1|$ is so small as to be of the order of C, the terms forming the summation will not vary very rapidly over the range of variation of Γ which is of interest.

We will thus write the circuit equation in the approximate form

$$E = \left[\frac{\Gamma^2\Gamma_1(E^2/\beta^2P)}{2(\Gamma_1^2 - \Gamma^2)} - \frac{j\Gamma^2}{\omega C_1}\right] i \tag{7.1}$$

Here there has been a simplification of notation. E is the z component of electric field, as in Chapter II, and is assumed to vary as $\exp(-\Gamma z)$. (E^2/β^2P) is taken to mean the value for the Γ_1 mode. It has been assumed that β_0^2 is small compared with $|\Gamma_1^2|$ and $|\Gamma^2|$, and β_0^2 has been neglected in comparison with these quantities.

Further, it has been pointed out that for slightly lossy circuits, (E^2/β^2P) will have only a small imaginary component, and we will assume as a valid approximation that (E^2/β^2P) is purely real. We cannot, however, safely assume that Γ_1 is purely imaginary, for a small real component of Γ_1 can affect the value of $\Gamma_1^2 - \Gamma^2$ greatly when Γ is nearly equal to Γ_1.

The first term on the right of (7.1) represents fields associated with the active mode of the circuit, which is nearly in synchronism with the electrons. We can think of these fields as summing up the effect of the electrons on the circuit over a long distance, propagated to the point under consideration.

The term $(-j\Gamma^2/\omega C_1)$ in (7.1) sums up the effect of all passive modes and of any active modes which are far out of synchronism with the electrons. It has been written in this form for a special purpose; the term will be regarded as constant over the range of Γ considered, and C_1 will be given a simple physical meaning.

This second term represents the field resulting from the local charge density, as opposed to that of the circuit wave which travels to the region from remote points. Let us rewrite (7.1) in terms of voltage and charge density

$$E = -\frac{\partial V}{\partial z} = \Gamma V \tag{7.2}$$

From the continuity equation

$$i = (j\omega/\Gamma)\rho \tag{2.18}$$

$$V = \left[\frac{j\omega\Gamma_1(E^2/\beta^2P)}{2(\Gamma_1^2 - \Gamma^2)} + \frac{1}{C_1}\right]\rho \tag{7.3}$$

We see that C_1 has the form of a capacitance per unit length. We can, for instance, redraw the transmission-line analogue of Fig. 2.3 as shown in Fig. 7.1. Here, the current I is still the line current; but the voltage V acting on the beam is the line voltage plus the drop across a capacitance of C_1 farads per meter.

Consider as an illustration the case of unattenuated waves for which

$$\Gamma_1 = j\beta_1 \tag{7.5}$$

$$\Gamma = j\beta \tag{7.6}$$

where β_1 and β are real. Then

$$V = \left[\frac{\omega\beta_1(E^2/\beta^2P)}{2(\beta_1^2 - \beta^2)} + \frac{1}{C_1}\right]\rho \tag{7.7}$$

In (7.7), the first term in the brackets represents the impedance presented to the beam by the "circuit"; that is, the ladder network of Figs. 2.3 and 7.1. The second term represents the additional impedance due to the capacitance C_1, which stands for the impedance of the nonsynchronous modes. We note that if $\beta < \beta_1$, that is, for a wave faster than the natural phase velocity of the circuit, the two terms on the right are of the same sign. This must mean that the "circuit" part of the impedance is capacitive. However, for $\beta > \beta_1$, that is, for a wave slower than the natural phase velocity, the first term is negative and the "circuit" part of the impedance is inductive. This is easily explained. For small values of β the wavelength of the impressed current is long, so that it flows into and out of the circuit at widely separated points. Between such points the long section of series inductance has a higher impedance than the shunt capacitance to ground; the capacitive effect predominates and the circuit impedance is capacitive. However, for large values of β the current flows into and out of the circuit at points close together. The short section of series inductance between such points provides a lower impedance path than does the shunt capacitance to ground; the inductive impedance predominates and the circuit impedance is inductive. Thus, for *fast* waves the circuit appears *capacitive* and for *slow* waves the circuit appears *inductive*.

Since we have justified the use of the methods of Chapter II within the limitations of certain assumptions, there is no reason why we should not proceed to use the same notation in the light of our fuller understanding. We can now, however, regard V not as a potential but merely as a convenient variable related to the field by (7.2).

From (2.18) and (7.3) we obtain

$$V = \left[\frac{\Gamma\Gamma_1(E^2/\beta^2P}{2(\Gamma_1^2 - \Gamma^2)} - \frac{j\Gamma}{\omega C_1}\right] i \tag{7.8}$$

We use this together with (2.22)

$$i = \frac{jI_0\beta_e\Gamma V}{2V_0(j\beta_e - \Gamma)^2} \tag{2.22}$$

We obtain the overall equation

$$1 = \frac{jI_0\beta_e\Gamma}{2V_0(j\beta_e - \Gamma)}\left[\frac{\Gamma\Gamma_1(E^2/\beta^2P)}{2(\Gamma_1 - \Gamma)} - \frac{j\Gamma}{\omega C_1}\right] \tag{7.9}$$

In terms of the gain parameter C, which was defined in Chapter II,

$$C^3 = (E^2/\beta^2P)(I_0/8V_0) \tag{2.43}$$

we can rewrite (7.8)

$$(j\beta_e - \Gamma)^2 = \frac{j2\beta_e\Gamma^2\Gamma_1C^3}{(\Gamma_1^2 - \Gamma^2)} + \frac{4\beta_e\Gamma^2C^3}{\omega C_1(E^2/\beta^2P)} \tag{7.10}$$

We will be interested in cases in which Γ and Γ_1 differ from β_e by a small amount only. Accordingly, we will write

$$-\Gamma = -j\beta_e + \beta_eC \tag{7.11}$$

$$-\Gamma_1 = -j\beta_e - j\beta_eCb - \beta_eCd \tag{7.12}$$

The propagation constant Γ describes propagation in the presence of electrons. A positive real value of δ means an increasing wave. A positive imaginary part means a wave traveling faster than the electrons.

The propagation constant Γ_1 refers to propagation in the circuit in the absence of electrons. A positive value of b means the electrons go faster than the undisturbed wave. A positive value d means that the wave is an attenuated wave which decreases as it travels.

If we use (7.11) and (7.12) in connection with (7.10) we obtain

$$\delta = \frac{[1 + C(2j\delta - C\delta^2)][1 + C(b - jd)]}{[-b + jd + j\delta + C(jbd - b^2/2 + d^2/2 + \delta^2/2)]} - \frac{4\beta_e[(1 + C(2j\delta - C\delta^2)]C}{\omega C_1(E^2/\beta^2P)} \tag{7.13}$$

We will now assume that $|\delta|$ is of the order of unity, that $|b|$ and $|d|$ range from zero to unity or a little larger, and that $C \ll 1$. We will then neglect the parentheses multiplied by C, obtaining

$$\delta = \frac{1}{(-b + jd + j\delta)} - 4QC \tag{7.14}$$

$$Q = \frac{\beta_e}{\omega C_1(E^2/\beta^2P)} \tag{7.15}$$

The quantity ωC_1 has the dimensions of admittance per unit length, β_e has the dimensions of (length)$^{-1}$ and $(E^2/\beta^2 P)$ has the dimensions of impedance. Thus, Q is a dimensionless parameter (the space-charge parameter) which may be thought of as relating to the impedance parameter $(E^2/\beta^2 P)$ associated with the synchronous mode the impedance $(\beta_e/\omega C_1)$, attributable to all modes but the synchronous mode.

At this point it is important to remember that there are not only two impedances, but two voltage components as well. Thus, in (7.8), the first term in the brackets times the current represents the "circuit voltage", which we may call V_c. The second term in the brackets represents the voltage due to space charge, the voltage across the capacitances C_1. The two terms in the brackets are in the same ratio as the two terms on the right of (7.14), which came from them. Thus, we can express the circuit component of voltage V_c in terms of the total voltage V acting on the beam either from (7.8) as

$$V_c = \left[1 - \frac{j2(\Gamma_1^2 - \Gamma^2)}{\omega C_1 \Gamma_1 (E^2/\beta^2 P)}\right]^{-1} V \tag{7.16}$$

or, alternatively, from (7.14) as

$$V_c = [1 - 4QC(-b + jd + j\delta)]^{-1} V \tag{7.17}$$

From Chapter II we have relations for the electron velocity (2.15) and electron convection current (2.22). If we make the same approximations which were made in obtaining (7.14), we have

$$(ju_0 C/\eta)v = \frac{V}{\delta} \tag{7.18}$$

$$(-2V_0 C^2/I)i = \frac{V}{\delta^2} \tag{7.19}$$

We should remember also that the variation of all quantities with z is as

$$e^{-j\beta_e z} e^{\beta_e C\delta z} \tag{7.20}$$

The relations (7.18)–(7.19) together with (2.36), which tells us that the characteristic impedance of the circuit changes little in the presence of electrons if C is small, sum up in terms of the more important parameters the linear operation of traveling-wave tubes in which C is small. The parameters are: the gain parameter C, relative electron velocity parameter b, circuit attenuation parameter d and space-charge parameter Q. In follow-

ing chapters, the practical importance of these parameters in the operation of traveling-wave tubes will be discussed.

There are other effects not encompassed by these equations. The effect of transverse electron motions is small in most tubes because of the high focusing fields employed; it will be discussed in a later chapter. The differences between a field theory in which different fields act on different electrons and the theory leading to (7.14)–(7.20), which apply accurately only when all electrons at a given z-position are acted on by the same field, will also be discussed.

CHAPTER VIII

THE NATURE OF THE WAVES

Synopsis of Chapter

IN THIS CHAPTER we shall discuss the effect of the various parameters on the rate of increase and velocity of propagation of the three forward waves. Problems involving boundary conditions will be deferred to later chapters.

The three parameters in which we are interested are those of (7.13), that is, b, the velocity parameter, d, the attenuation parameter and QC, the space-charge parameter. The fraction by which the electron velocity is greater than the phase velocity for the circuit in the absence of electrons is bC. The circuit attenuation is 54.6 dC db/wavelength. Q is a factor depending on the circuit impedance and geometry and on the beam diameter. For a helically conducting sheet of radius a and a hollow beam of radius a_1, Q can be obtained from Fig. 8.12.

The three forward waves vary with distance as

$$e^{-j\beta_e(1-yC)z}e^{\beta_e xCz}$$

$$\beta_e = \frac{\omega}{u_0}$$

Thus, a positive value of y means a wave which travels faster than the electrons, and a positive value of x means an increasing wave. The gain in db per wavelength of the increasing waves is BC, and B is defined by (8.9).

Figure 8.1 shows x and y for the three forward waves for a lossless circuit ($d = 0$). The increasing wave is described by x_1, y_1. The gain is a maximum when the electron velocity is equal to the velocity of the undisturbed wave, or, when $b = 0$. For large positive values of b (electrons much faster than undisturbed wave), there is no increasing wave. However, there is an increasing wave for all negative values of b (all low velocities). For the increasing wave, y_1 is negative; thus, the increasing wave travels more slowly than the electrons, *even when the electrons travel more slowly than the circuit wave in the absence of electrons.* For the range of b for which there is an increasing wave, there is also an attenuated wave, described by $x_2 = -x_1$ and $y_2 = y_1$. There is also an unattenuated wave described by $y_3(x_3 = 0)$.

For very large positive and negative values of b, the velocity of two of the waves approaches the electron velocity (y approaches zero) and the

velocity of the third wave approaches the velocity of the circuit wave in the absence of electrons (y approaches minus b). For large negative values of b, x_1, y_1 and x_2, y_2 become the "electron" waves and y_3 becomes the "circuit" wave. For large values of b, y_1 and y_3 become the "electron" waves and y_2 becomes the "circuit" wave. The "circuit" wave is essentially the wave in the absence of electrons, modified slightly by the presence of a non-synchronous electron stream. The "electron waves" represent the motion of "bunches" along the electron stream, slightly affected by the presence of the circuit.

Figures 8.2 and 8.3 indicate the effect of loss. Loss decreases the gain of the increasing wave, adds to the attenuation of the decreasing wave and adds attenuation to the wave which was unattenuated in the lossless case. For large positive and negative values of b, the attenuation of the circuit wave (given by x_3 for negative values of b and x_2 for positive values of b) approaches the attenuation in the absence of electrons.

Figure 8.4 shows B, the gain of the increasing wave in db per wavelength per unit C. Figure 8.5 shows, for $b = 0$, how B varies with d. The dashed line shows a common approximation: that the gain of the increasing wave is reduced by $\frac{1}{3}$ of the circuit loss. Figure 8.6 shows how, for $b = 0$, x_1, x_2 and x_3 vary with d. We see that, for large values of d, the wave described by x_2 has almost the same attenuation as the wave on the circuit in the absence of electrons.

Figures 8.7–8.9 show x, y for the three waves with no loss ($d = 0$) but with a-c space charge taken into account ($QC \neq 0$). The immediately striking feature is that there is now a minimum value of b below which there is no increasing wave.

We further note that, for large negative and positive values of b, y for the electron waves approaches $\pm 2\sqrt{QC}$. In these ranges of b the electron waves are dependent on the electron inertia and the field produced by a-c space charge, and have nothing to do with the active mode of the circuit.

As QC is made larger, the value of b for which the gain of the increasing wave is a maximum increases. Now, C is proportional to the cube root of current. Thus, as current is increased, the voltage for maximum gain of the increasing wave increases. An increase in optimum operating voltage with an increase in current is observed in some tubes, and this is at least partly explained by these curves.* There is also some decrease in the maximum value of x_1 and hence of B as QC is increased. This is shown more clearly in Fig. 8.10.

If x and B remained constant when the current is varied, then the gain per wavelength would rise as C, or, as the $\frac{1}{3}$ power of current. However,

* Other factors include a possible lowering of electron speed because of d-c space charge, and boundary condition effects.

we see from Fig. 8.10 that B falls as QC is increased. The gain per wavelength varies as BC and, because Q is constant for a given tube, it varies as BQC. In Fig. 8.11, BQC, which is proportional to the gain per wavelength of the increasing wave, is plotted vs QC, which is proportional to the $\frac{1}{3}$ power of current. For very small values of current (small values of QC), the gain per wavelength is proportional to the $\frac{1}{3}$ power of current. For larger values of QC, the gain per wavelength becomes proportional to the $\frac{1}{4}$ power of current.

It would be difficult to present curves covering the simultaneous effect of loss (d) and space charge (QC). As a sort of substitute, Figs. 8.13 and 8.14 show $\partial x_1/\partial d$ for $d = 0$ and b chosen to maximize x_1, and $\partial x_1/\partial(QC)$ for $QC = 0$ and $b = 0$. We see from 8.13 that, while for small values of QC the gain of the increasing wave is reduced by $\frac{1}{3}$ of the circuit loss, for large values of QC the gain of the increasing wave is reduced by $\frac{1}{2}$ of the circuit loss.

8.1 Effect of Varying the Electron Velocity

Consider equation (7.13) in case $d = 0$ (no attenuation) and $Q = 0$ (neglect of space-charge). We then have

$$\delta^2(\delta + jb) = -j \tag{8.1}$$

Here we will remember that

$$\beta_e = \omega/u_0 \tag{8.2}$$

$$-\Gamma_1 = -j\beta_e(1 + Cb) = -j\omega/v_1 \tag{8.3}$$

Here v_1 is the phase velocity of the wave in the absence of electrons, and u_0 is the electron speed. We see that

$$u_0 = (1 + Cb)v_1 \tag{8.4}$$

Thus, $(1 + Cb)$ is the ratio of the electron velocity to the velocity of the *undisturbed wave*, that is, the wave in the absence of electrons. Hence, b is a measure of velocity difference between electrons and undisturbed wave. For $b > 0$, the electrons go faster than the undisturbed wave; for $b < 0$ the electrons go slower than the undisturbed wave. For $b = 0$ the electrons have the same speed as the undisturbed wave.

If $b = 0$, (8.1) becomes

$$\delta^3 = -j \tag{8.5}$$

which we obtained in Chapter II.

In dealing with (8.1), let

$$\delta = x + jy$$

The meaning of this will be clear when we remember that, in the presence of electrons, quantities vary with z as (from (7.10))

$$\begin{aligned} & e^{-j\beta_e(1+jC\delta)z} \\ &= e^{-j\beta_e(1-Cy)z} e^{\beta_e Cxz} \end{aligned} \tag{8.6}$$

If v is the phase velocity in the presence of electrons, we have

$$\omega/v = (\omega/u_0)(1 - Cy) \tag{8.7}$$

If $Cy \ll 1$, very nearly

$$v = u_0(1 + Cy) \tag{8.8}$$

In other words, if $y > 0$, the wave travels faster than the electrons; if $y < 0$ the wave travels more slowly than the electrons.

From (8.6) we see that, if $x > 0$, the wave increases as it travels and if $x < 0$ the wave decreases as it travels. In Chapter II we expressed the gain of the increasing wave as

$$BCN \text{ db}$$

where N is the number of wavelengths. We see that

$$\begin{aligned} B &= 20(2\pi)(\log_{10} e)x \\ B &= 54.5x \end{aligned} \tag{8.9}$$

In terms of x and y, (8.1) becomes

$$(x^2 - y^2)(y + b) + 2x^2y + 1 = 0 \tag{8.10}$$

$$x(x^2 - 3y^2 - 2yb) = 0 \tag{8.11}$$

We see that (8.11) yields two kinds of roots: those corresponding to unattenuated waves, for which $x = 0$ and those for which

$$x^2 = 3y^2 + 2yb \tag{8.12}$$

If $x = 0$, from (8.10)

$$\begin{aligned} y^2(y + b) &= 1 \\ b &= -y + 1/y^2 \end{aligned} \tag{8.13}$$

If we assume values of y ranging from perhaps $+4$ to -4 we can find the corresponding values of b from (8.13), and plot out y vs b for these unattenuated waves.

For the other waves, we substitute (8.12) into (8.10) and obtain

$$2yb^2 + 8y^2b + 8y^3 + 1 = 0 \tag{8.14}$$

This equation is a quadratic in b, and, by assigning various values of y, we can solve for b. We can then obtain x from (8.12).

In this fashion we can construct curves of x and y vs b. Such curves are shown in Fig. 8.1.

We see that for

$$b < (3/2)(2)^{1/3}$$

there are two waves for which $x \neq 0$ and one unattenuated wave. The increasing and decreasing waves ($x \neq 0$) have equal and opposite values of x, and since for them $y < 1$, they travel more slowly than the electrons, *even when the electrons travel more slowly than the undisturbed wave.* It can be

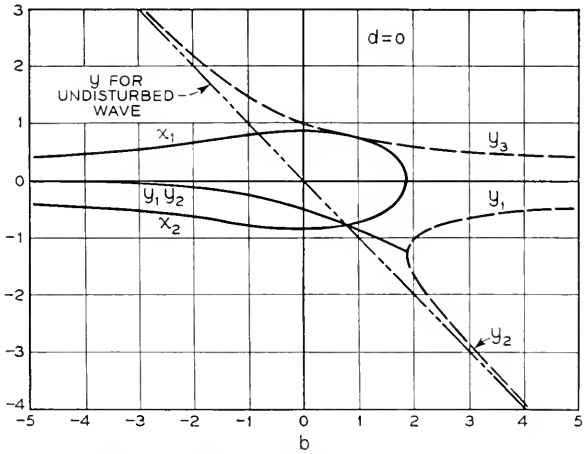

Fig. 8.1—The three waves vary with distance as exp $(-j\beta_e + j\beta_e Cy + \beta_e Cx)z$. Here the x's and y's for the three waves are shown vs the velocity parameter b for no attenuation ($d = 0$) and no space charge ($QC = 0$).

shown that the electrons must travel faster than the increasing wave in order to give energy to it.

For $b > (3/2)(2)^{1/3}$, there are 3 unattenuated waves: two travel faster than the electrons and one more slowly.

For large positive or negative values of b, two waves have nearly the electron speed ($| y |$ small) and one wave travels with the speed of the undisturbed wave. We measure velocity with respect to electron velocity. Thus, if we assigned a parameter y to describe the velocity of the undisturbed wave relative to the electron velocity, it would vary as the 45° line in Fig. 8.1.

The data expressed in Fig. 8.1 give the variation of gain per wavelength of the undisturbed wave with electron velocity, and are also useful in fitting

boundary conditions; for this we need to know the three x's and the three y's.

In a tube in which the total gain is large, a change in b of $\pm$ 1 about $b = 0$ can make a change of several db in gain. Such a change means a difference between phase velocity of the undisturbed wave, v_1, and electron velocity u_0 by a fraction approximately $\pm C$. Hence, the allowable difference between phase velocity v_1 of the undisturbed wave, which is a function of frequency, and electron velocity, which is not, is of the order of C.

8.2 Effect of Attenuation

If we say that $d \neq 0$ but has some small positive value, we mean that the circuit is lossy, and in the absence of electrons the voltage decays with distance as

$$e^{-\beta_e C d}$$

Hence, the loss L in db/wavelength is

$$\begin{aligned} L &= 20(2\pi)(\log_{10} e) C d \\ L &= 54.5 C d \text{ db/wavelength} \end{aligned} \tag{8.15}$$

or

$$d = .01836\ (L/C) \tag{8.16}$$

For instance, for $C = .025$, $d = 1$ means a loss of 1.36 db wavelength.

If we assume $d \neq 0$ we obtain the equations

$$(x^2 - y^2)(y + b) + 2xy(x + d) + 1 = 0 \tag{8.17}$$

$$(x^2 - y^2)(x + d) - 2xy(y + b) = 0 \tag{8.18}$$

The equations have been solved numerically for $d = .5$ and $d = 1$, and the curves which were obtained are shown in Figs. 8.2 and 8.3. We see that for a circuit with attenuation there is an increasing wave for all values of b (electron velocity). The velocity parameters y_1 and y_2 are now distinct for all values of b.

We see that the maximum value of x_1 decreases as loss is increased. This can be brought out more clearly by showing x_1 vs b on an expanded scale. It is perhaps more convenient to plot B, the db gain per wavelength per unit C, vs b, and this has been done for various values of d in Fig. 8.4.

We see that for small values of d the maximum value of x_1 occurs very near to $b = 0$. If we let $b = 0$ in (8.17) and (8.18) we obtain

$$y(x^2 - y^2) + 2xy(x + d) + 1 = 0 \tag{8.19}$$

$$(x^2 - y^2)(x + d) - 2xy^2 = 0 \tag{8.20}$$

We can rewrite (8.20) in the form

$$y = \pm x\left(\frac{1 + d/x}{3 + d/x}\right)^{1/2} \tag{8.21}$$

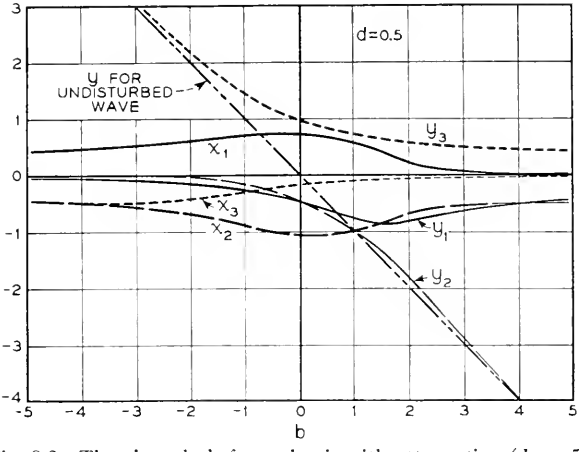

Fig. 8.2—The x's and y's for a circuit with attenuation ($d = .5$).

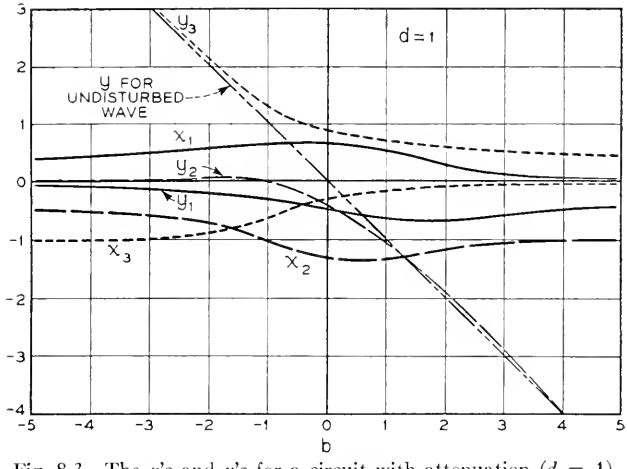

Fig. 8.3—The x's and y's for a circuit with attenuation ($d = 1$).

If we substitute this into (8.19) we can solve for x in terms of the parameter d/x

$$x = \mp\left[\frac{\left(\dfrac{3 + d/x}{1 + d/x}\right)^{1/2}}{2\left(\dfrac{1}{3 + d/x} + 1 + d/x\right)}\right]^{1/3} \tag{8.22}$$

Here we take both upper signs or both lower signs in (8.21) and (8.22).

If we assume $d/x \ll 1$ and expand, keeping no powers of d/x higher than the first, we obtain

$$x = \mp (\sqrt{3}/2)(1 - (1/3(d/x)) \tag{8.23}$$

The plus sign will give x_1, which is the x for the increasing wave. Let x_{10} be the value of x_1 for $d = 0$ (no loss).

$$x_{10} = \sqrt{3}/2 \tag{8.24}$$

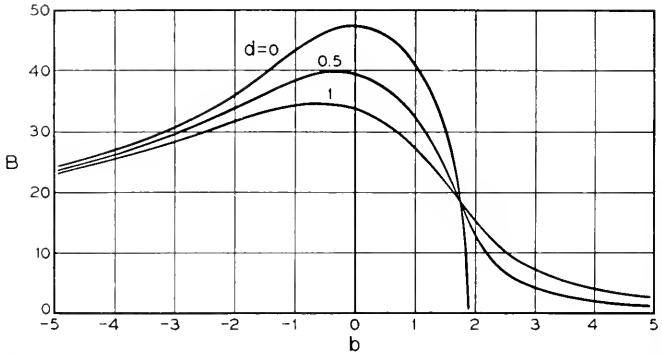

Fig. 8.4—The gain of the increasing wave is BCN db, where N is the number of wavelengths.

Then for small values of d

$$x_1 = x_{10}(1 - (1/3)(d/x_{10}))$$

$$x_1 = x_{10} - 1/3d \tag{8.25}$$

This says that, for small losses, the reduction of gain *of the increasing wave* from the gain in db for zero loss is $\frac{1}{3}$ of the circuit attenuation in db. The reduction of *net gain*, which will be greater, can be obtained only by matching boundary conditions in the presence of loss (see Chapter IX).

In Fig. 8.5, $B = 54.6\, x_1$ has been plotted vs d from (8.22). The straight line is for $x_{10} = d/3$.

In Fig. 8.6, $-x_1$, x_2 and x_3 have been plotted vs d for a large range in d. As the circuit is made very lossy, the waves which for no loss are unattenuated and increasing turn into a pair of waves with equal and opposite small attenuations. These waves will be essentially disturbances in the electron stream, or space-charge waves. The original decreasing wave turns into a wave which has the attenuation of the circuit, and is accompanied by small disturbances in the electron stream.

8.3 Space-Charge Effects

Suppose that we let d, the attenuation parameter, be zero, but consider cases in which the space-charge parameter QC is not zero. We then obtain

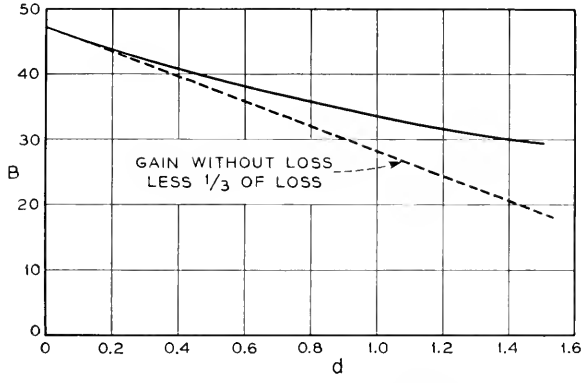

Fig. 8.5—For $b = 0$, that is, for electrons with a velocity equal to the circuit phase velocity, the gain factor B falls as the attenuation parameter d is increased. For small values of d, the gain is reduced by $\frac{1}{3}$ of the circuit loss.

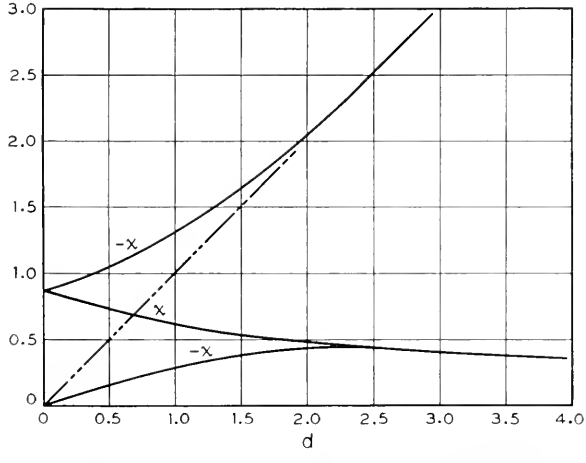

Fig. 8.6—How the three x's vary for $b = 0$ and for large losses.

the equations

$$(x^2 - y^2)(b + y) + 2x^2y + 4QC(b + y) + 1 = 0 \qquad (8.26)$$

$$x[(x^2 - y^2) - 2y(y + b) + QC] = 0 \qquad (8.27)$$

Solutions of this have been found by numerical methods for $QC = .25$, .5 and 1; these are shown in Figs. 8.7–8.9.

We see at once that the electron velocity for maximum gain shifts markedly as QC is increased. Hence, the region around $b = 0$ is not in this case worthy of a separate investigation.

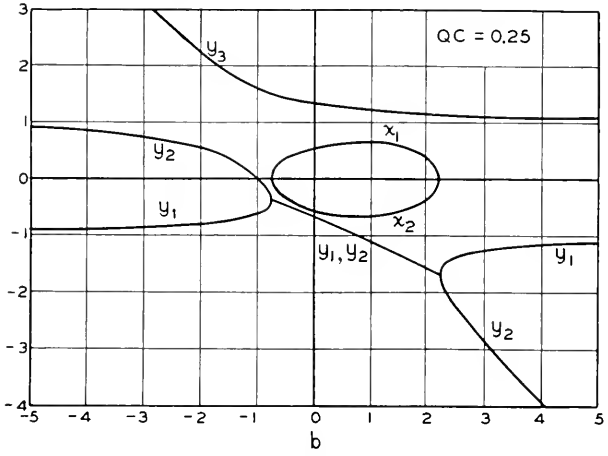

Fig. 8.7—The x's and y's for the three waves with zero loss ($d = 0$) but with space charge ($QC = .25$).

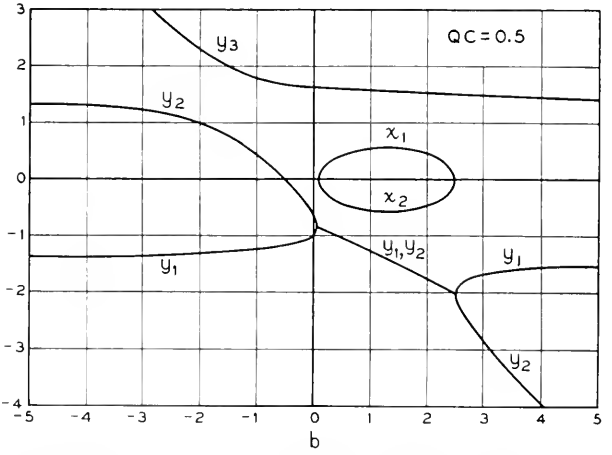

Fig. 8.8 The x's and y's with greater space charge ($QC = .5$).

It is interesting to plot the maximum value of x_1 vs. the parameter QC. This has, in effect, been done in Fig. 8.10, which shows B, the gain in db per wavelength per unit C, vs. QC.

We can obtain a curve proportional to db per wavelength by plotting BQC vs. QC. (Q is independent of current.) This has been done in Fig. 8.11. For $QC < 0.025$, the gain in db per wavelength varies linearly with

QC. Chu and Rydbeck found that under certain conditions gain varies approximately as the $\frac{1}{4}$ power of the current. This would mean a slope of $\frac{3}{4}$ on Fig. 8.11. A $\frac{3}{4}$ power dashed line is shown in Fig. 8.11; it fits the upper part of the curve approximately.

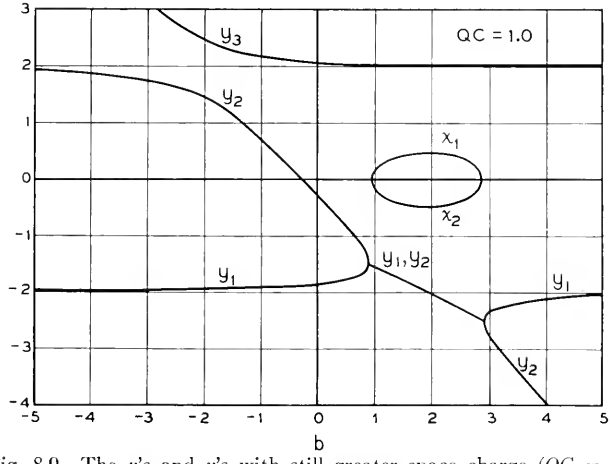

Fig. 8.9—The x's and y's with still greater space charge ($QC = 1$).

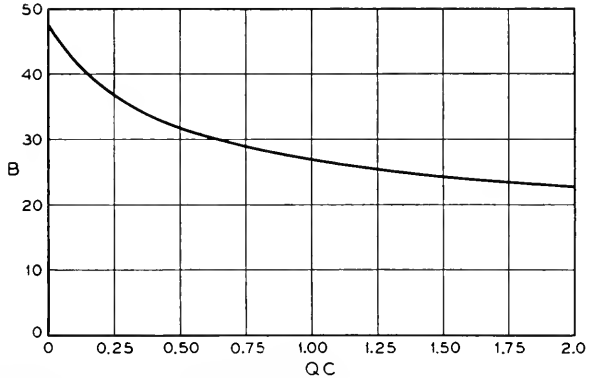

Fig. 8.10—How the gain factor B decreases as QC is increased, for the value of b which gives a maximum value of x_1.

If we examine Figs. 8.7–8.9 we find that for large and small values of b there are, as in other cases, a circuit wave, for which y is nearly equal to $-b$, and two space-charge waves. For these, however, y does not approach zero.

Let us consider equation (7.13). If b is large, the first term on the right becomes small, and we have approximately

$$\delta = \pm j2\sqrt{QC} \tag{8.28}$$

These waves correspond to the space-charge waves of Hahn and Ramo, and are quite independent of the circuit impedance, which appears in (8.28) merely as an arbitrary parameter defining the units in which δ is measured. Equation (8.28) also describes the disturbance we would get if we shorted out the circuit by some means, as by adding excessive loss.

Practically, we need an estimate of the value of Q for some typical circuit. In Appendix IV an estimate is made on the following basis: The helix

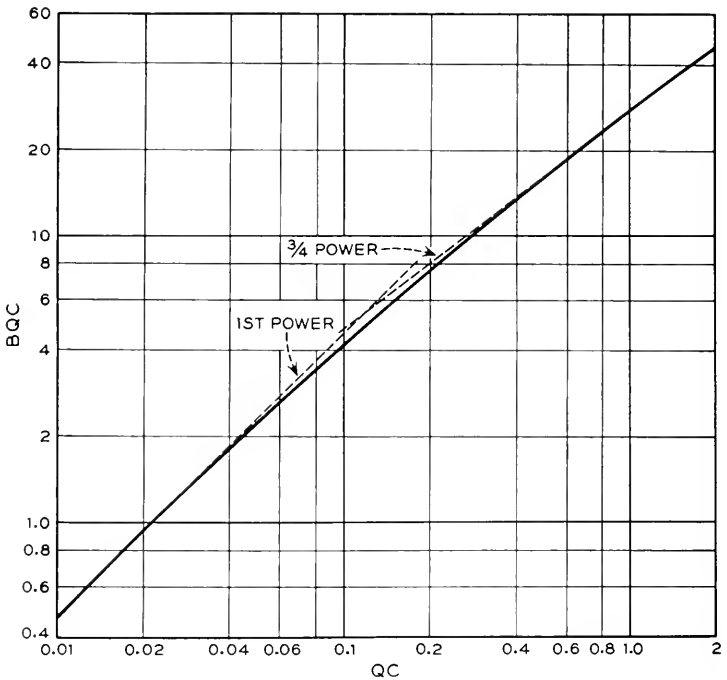

Fig. 8.11—The variation of a quantity proportional to the cube of the gain of the increasing wave (ordinate) with a quantity proportional to current (abscissa). For very small currents, the gain of the increasing wave is proportional to the $\frac{1}{3}$ power of current, for large currents to the $\frac{1}{4}$ power of current.

of radius a is replaced by a conducting cylinder of the same radius, a thin cylinder of convection current of radius a_1 and current of $i \exp(-j\beta z)$ is assumed, and the field is calculated and identified with the second term on the right of (7.1). R. C. Fletcher has obtained a more accurate value of Q by a rigorous method. His work is reproduced in Appendix VI, and in Fig. 1 of that appendix, Fletcher's value of Q is compared with the approximate value of Appendix IV.

In Fig. 8.12, the value $Q(\beta\ \gamma)^2$ of Appendix IV is plotted vs. γa for $a_1/a = .9,\ .8,\ .7$. For $a_1/a = 1$, $Q = 0$. In a typical 4,000 mc traveling-wave

tube, $\gamma a = 2.8$ and C is about .025. Thus, if we take the effective beam radius as .5 times the helix radius, $Q = 5.6$ and $QC = .14$.

We note from (7.14) that Q is the ratio of a capacitive impedance to $(E^2/\beta^2 P)$. In obtaining the curves of Fig. 8.12, the value of $(E^2/\beta^2 P)$ for a helically conducting sheet was assumed. This is given by (3.8) and (3.9). If $(E^2/\beta^2 P)$ is different for the circuit actually used, and it is somewhat different, even for an actual helix, Q from Fig. 8.12 should be multiplied by $(E^2/\beta^2 P)$ for the helically conducting sheet, from (3.8) and (3.9), and divided by the value of $(E^2/\beta^2 P)$ for the circuit used.

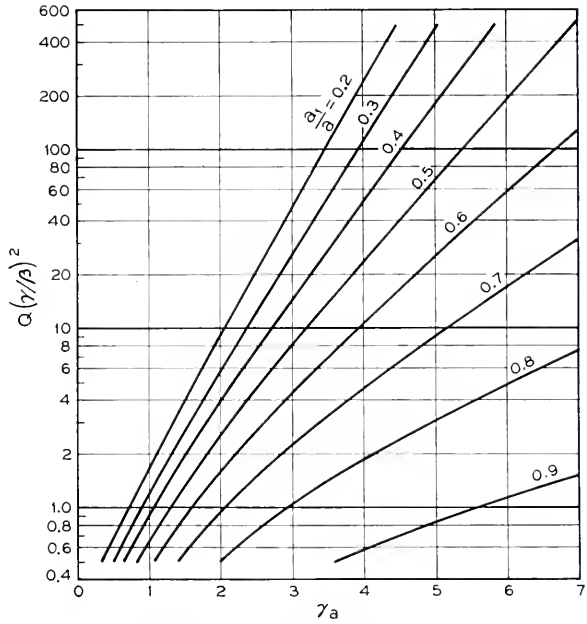

Fig. 8.12—Curves for obtaining Q for a helically conducting sheet and a hollow beam. The radius of the helically conducting sheet is a and that of the beam is a_1.

8.4 Differential Relations

It would be onerous to construct curves giving δ as a function of b for many values of attenuation and space charge. In some cases, however, useful information may be obtained by considering the effect of adding a small amount of attenuation when QC is large, or of seeing the effect of space charge when QC is small but the attenuation is large. We start with (7.13)

$$\delta^2 = \frac{1}{(-b + jd + j\delta)} - 4QC \tag{7.13}$$

Let us first differentiate (7.13) with respect to δ and d

$$2\delta\, d\delta = \frac{-j\, dd - j\, d\delta}{(-b + j\, d + j\delta)^2} \tag{8.29}$$

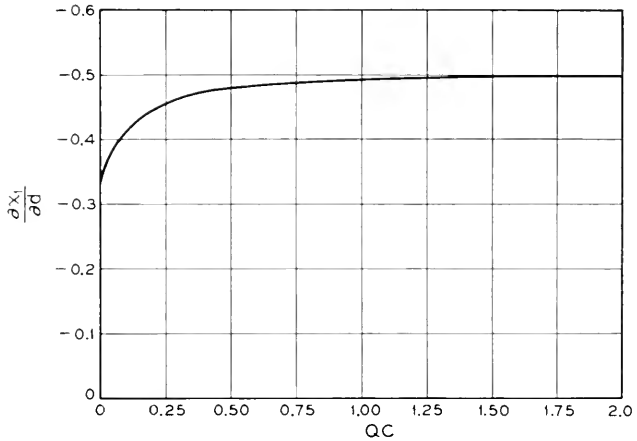

Fig. 8.13—A curve giving the rate of change of x_1 with attenuation parameter d for $d = 0$ and for various values of the space-charge parameter QC. For small values of QC the gain of the increasing wave is reduced by $\frac{1}{3}$ of the circuit loss; for large values of QC the gain of the increasing wave is reduced by $\frac{1}{2}$ of the circuit loss.

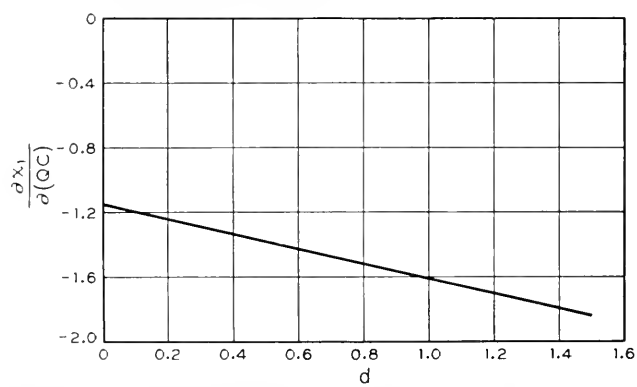

Fig. 8.14—A curve showing the variation of x_1 with QC for $QC = 0$ and for various values of the attenuation parameter d.

By using (7.13) we obtain

$$d\delta = \left(\frac{-j2\delta}{(\delta^2 + 4QC)^2} - 1\right)^{-1} dd \tag{8.30}$$

If we allow d to be small, we can use the values of δ of Figs. 8.7–8.9 to plot the quantity

$$\mathrm{Re}(d\delta_1/dd) = dx_1/dd \tag{8.31}$$

vs. QC. In Fig. 8.13, this has been done for b chosen to make x_1 a maximum. We see that a small loss dd causes more reduction of gain as QC is increased (more space charge).

Let us now differentiate (7.13) with respect to QC

$$2\delta\, d\delta = \frac{-j\, d\delta}{(-b + j\, d + j\delta)^2} - 4\, d(QC) \tag{8.32}$$

By using (7.13) with $QC = 0$ we obtain

$$d\delta = \left(\frac{-4}{2\delta + j\delta^4}\right) d(QC) \tag{8.33}$$

In Fig. 8.14, $dx/d(QC)$ has been plotted vs. d for $b = 0$.

We see that the reduction of gain for a small amount of space charge becomes greater, the greater the loss is increased (d increased).

Both Fig. 8.13 and Fig. 8.14 indicate that for large values of QC or d the gain will be overestimated if space charge (QC) and loss (d) are considered separately.

CHAPTER IX

DISCONTINUITIES

Synopsis of Chapter

WE WANT TO KNOW the overall gain of traveling-wave tubes. So far, we have evaluated only the gain of the increasing wave, and we must find out how strong an increasing wave is set up when a voltage is applied to the circuit.

Beyond this, we may wish for some reason to break the circuit up into several sections having different parameters. For instance, it is desirable that a traveling-wave tube have more loss in the backward direction than it has gain in the forward direction. If this is not so, small mismatches will result either in oscillation or at least in the gain fluctuating violently with frequency. We have already seen in Chapter VIII the effect of a uniform loss in reducing the gain of the increasing wave. We need to know also the overall effect of short sections of loss in order to know how loss may best be introduced.

Such problems are treated in this chapter by matching boundary conditions at the points of discontinuity. It is assumed that there is no reflected wave at the discontinuity. This will be very nearly so, because the characteristic impedances of the waves differ little over the range of loss and velocity considered. Thus, the total voltages, a-c convection currents and the a-c velocities on the two sides of the point of discontinuity are set equal.

For instance, at the beginning of the circuit, where the unmodulated electron stream enters, the total a-c velocity and the total a-c convection current—that is, the sums of the convection currents and the velocities for the three waves—are set equal to zero, and the sum of the voltages for the three waves is set equal to the applied voltage.

For the case of no loss ($d = 0$) and an electron velocity equal to circuit phase velocity ($b = 0$) we find that the three waves are set up with equal voltages, each $\frac{1}{3}$ of the applied voltage. The voltage along the circuit will then be the sum of the voltages of the three waves, and the way in which the magnitude of this sum varies with distance along the circuit is shown in Fig. 9.1. Here CN measures distance from the beginning of the circuit and the amplitude relative to the applied voltage is measured in db.

The dashed curve represents the voltage of the increasing wave alone.

For large values of CN corresponding to large gains, the increasing wave predominates and we can neglect the effect of the other waves. This leads to the gain expression

$$G = A + BCN \text{ db}$$

Here BCN is the gain in db of the increasing wave and A measures its initial level with respect to the applied voltage.

In Fig. 9.2, A is plotted vs. b for several values of the loss parameter d. The fact that A goes to ∞ for $d = 0$ as b approaches $(3/2)\ (2)^{1/3}$ does not imply an infinite gain for, at this value of b, the gain of the increasing wave approaches zero and the voltage of the decreasing wave approaches the negative of that for the increasing wave.

Figure 9.3 shows how A varies with d for $b = 0$. Figure 9.4 shows how A varies with QC for $d = 0$ and for b chosen to give a maximum value of B (the greatest gain of the increasing wave).

Suppose that for $b = QC = 0$ the loss parameter is suddenly changed from zero to some finite value d. Suppose also that the increasing wave is very large compared with the other waves reaching the discontinuity. We can then calculate the ratio of the increasing wave just beyond the discontinuity to the increasing wave reaching the discontinuity. The solid line of Fig. 9.5 shows this ratio expressed in decibels. We see that the voltage of the increasing wave excited in the lossy section is less than the voltage of the incident increasing wave.

Now, suppose the waves travel on in the lossy section until the increasing wave again predominates. If the circuit is then made suddenly lossless, we find that the increasing wave excited in this lossless section will have a greater voltage than the increasing wave incident from the lossy section, as shown by the dashed curve of Fig. 9.5. This increase is almost as great as the loss in entering the lossy section. Imagine a tube with a long lossless section, a long lossy section and another long lossless section. We see that the gain of this tube will be less than that of a lossless tube of the same total length by about the reduction of the gain of the increasing wave in lossy section.

Suppose that the electromagnetic energy of the circuit is suddenly absorbed at a distance beyond the input measured by CN. This might be done by severing a helix and terminating the ends. The a-c velocity and convection current will be unaffected in passing the discontinuity, but the circuit voltage drops to zero. For $d = b = QC = 0$, Fig. 9.6 shows the ratio of V_1, the amplitude of the increasing wave beyond the break, to V, the amplitude the increasing wave would have had if there were no break. We see that for CN greater than about 0.2 the loss due to the break is not

serious. For CN large (the break far from the input) the loss approaches 3.52 db.

Beyond such a break, the total voltage increases with CN as shown in Fig. 9.7, and from $CN = 0.2$ the circuit voltage is very nearly equal to the voltage of the increasing wave.

Often, for practical reasons loss is introduced over a considerable distance, sometimes by putting lossy material near to a helix. Suppose we use CN computed as if for a lossless section of circuit as a measure of length of the lossy section, and assume that the loss is great enough so that the circuit voltage (as opposed to that produced by space charge) can be taken as zero. Such a lossy section acts as a drift space. Suppose that an increasing wave only reaches this lossy section. The amplitude of the increasing wave excited beyond the lossy section in db with respect to the amplitude of the increasing wave reaching the lossy section is shown vs. CN, which measures the length of the lossy section, in Fig. 9.8.

9.1 General Boundary Conditions

We have already assumed that C is small, and when this is so the characteristic impedance of the various waves is near to the circuit characteristic impedance K. We will neglect any reflections caused by differences among the characteristic impedances of the various waves.

We will consider cases in which the circuit is terminated in the $+z$ direction, so as to give no backward wave. We will then be concerned with the 3 forward waves, for which δ has the values δ_1, δ_2, δ_3 and the waves represented by these values of δ have voltages V_1, V_2, V_3, electron velocities v_1, v_2, v_3 and convection currents i_1, i_2, i_3.

Let V, v, i be the total voltage, velocity and convection current at $z = 0$. Then we have

$$V_1 + V_2 + V_3 = V \tag{9.1}$$

and from (7.15) and (7.16),

$$\frac{V_1}{\delta_1} + \frac{V_2}{\delta_2} + \frac{V_3}{\delta_3} = (ju_0C/\eta)v \tag{9.2}$$

$$\frac{V_1}{\delta_1^2} + \frac{V_2}{\delta_2^2} + \frac{V_3}{\delta_3^2} = (-2V_0C^2/I_0)i \tag{9.3}$$

These equations yield, when solved,

$$V_1 = [V - (\delta_2 + \delta_3)(ju_0C/\eta)v + \delta_2\delta_3(-2V_0C^2/I_0)i] \\ [(1 - \delta_2\delta_1)(1 - \delta_3\delta_1)]^{-1} \tag{9.4}$$

We can obtain the corresponding expressions for V_2 and V_3 simply by inter-

changing subscripts; to obtain V_2, for instance, we substitute subscript 2 for 1 and subscript 1 for 2 in (9.4).

9.2 Lossless Helix, Synchronous Velocity, No Space Change

Suppose we consider the case in which $b = d = Q = 0$, so that we have the values of δ obtained in Chapter II

$$\begin{aligned} \delta_1 &= e^{-j\pi/6} = \sqrt{3}/2 - j1/2 \\ \delta_2 &= e^{-j5\pi/6} = -\sqrt{3}/2 - j1/2 \\ \delta_3 &= e^{j\pi/2} = j \end{aligned} \tag{9.5}$$

Suppose we inject an unmodulated electron stream into the helix and apply a voltage V. The obvious thing is to say that, at $z = 0$, $v = i = 0$. It is not quite clear, however, that $v = 0$ at $z = 0$ (the beginning of the circuit). Whether or not there is a stray field, which will give an initial velocity modulation, depends on the type of circuit. Two things are true, however. For the small values of C usually encountered such a velocity modulation constitutes a small effect. Also, the fields of the first part of the helix act essentially to velocity modulate the electron stream, and hence a neglect of any small initial velocity modulation will be about equivalent to a small displacement of the origin.

If, then, we let $v = i = 0$ and use (9.4) we obtain

$$V_1 = V[(1 - \delta_2/\delta_1)(1 - \delta_3/\delta_1)]^{-1} \tag{9.6}$$

$$V_1 = V/3 \tag{9.7}$$

Similarly, we find that

$$V_2 = V_3 = V/3 \tag{9.8}$$

We have used V to denote the voltage at $z = 0$. Let V_z be the voltage at z. We have

$$\begin{aligned} V_z &= (V/3)e^{-j\beta_e z}(e^{j(1/2)\beta_e Cz + (\sqrt{3}/2)\beta_e Cz} + e^{j(1/2)\beta_e Cz - (\sqrt{3}/2)\beta_e Cz} + e^{j\beta_e Cz}) \\ V_z &= (V/3)e^{-j\beta_e(1-C)z}\,(1 + 2\cos((\sqrt{3}/2)\beta_e Cz)e^{-j(3/2)\beta_e Cz}) \end{aligned} \tag{9.9}$$

From this we obtain

$$\begin{aligned} |V_z/V|^2 = (1/9)[1 &+ 4\cos^2(\sqrt{3}/2)\beta_e Cz \\ &+ 4\cos(3/2)\beta_e Cz\cos(\sqrt{3}/2)\beta_e Cz] \end{aligned} \tag{9.10}$$

We can express gain in db as $10\log_{10}|V_z/V|^2$, and, in Fig. 9.1, gain in db is plotted vs CN, where N is the number of cycles.

We see that initially the voltage does not change with distance. This is natural, because the electron stream initially has no convection current,

and hence cannot act on the circuit until it becomes bunched. Finally, of course, the increasing wave must predominate over the other two, and the slope of the line must be

$$B = 47.3/CN \tag{9.11}$$

The dashed line represents the increasing wave, which starts at $V_2/V = \frac{1}{3}$ (−9.54 db) and has the slope specified by (9.11). Thus, if we write for the increasing wave that gain G is

$$G = A + BCN \text{ db} \tag{9.12}$$

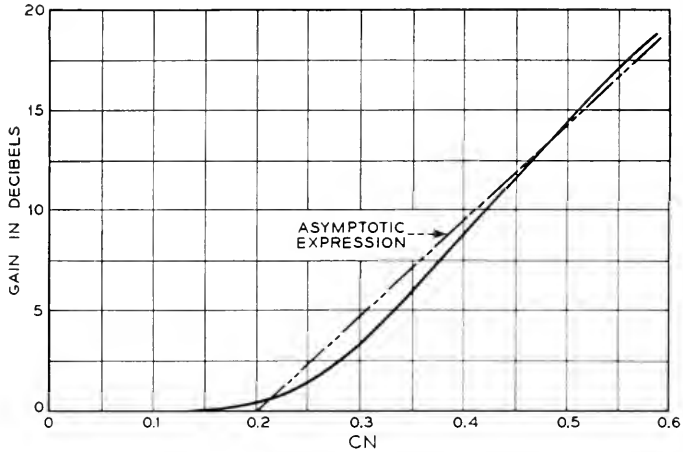

Fig. 9.1—How the signal level varies along a traveling-wave tube for the special case of zero loss and space charge and an electron velocity equal to the circuit phase velocity (solid curve). The dashed curve is the level of the increasing wave alone, which starts off with $\frac{1}{3}$ of the applied voltage, or at −9.54 *db*.

This is an asymptotic expression for the total voltage at large values of z, where $|V_1| \gg |V_2|, |V_3|$, and for $b = d = Q = 0$

$$\begin{aligned} A &= -9.54 \text{ db} \\ B &= 47.3 \end{aligned} \tag{9.13}$$

We see that (9.11) is pretty good for $CN > .4$, and not too bad for $CN > .2$.

9.3 Loss in Helix

In Chapter VIII, curves were given for δ_1, δ_2, δ_3 vs. b for $QC = 0$ and for d, the loss parameter, equal to 0, 0.5 and 1. From the data from which these curves were derived one can calculate the initial loss parameter by means of (9.6)

$$A = 20 \log_{10} |V_1/V| \tag{9.14}$$

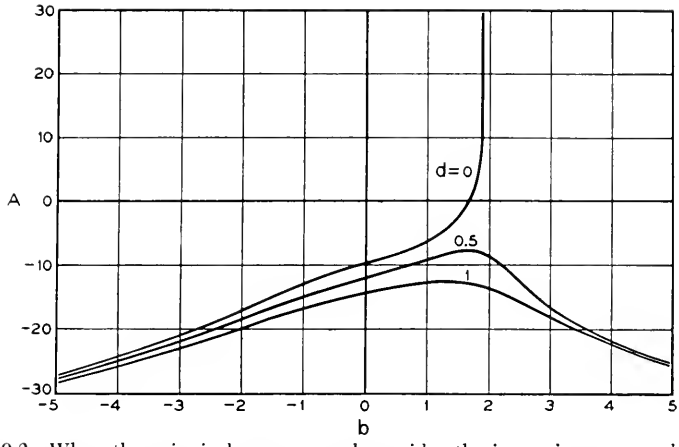

Fig. 9.2—When the gain is large we need consider the increasing wave only. Using this approximation, the gain in db is $A + BCN$ db. Here A is shown vs the velocity parameter b, several values of the attenuation parameter d, for no space charge ($QC = 0$).

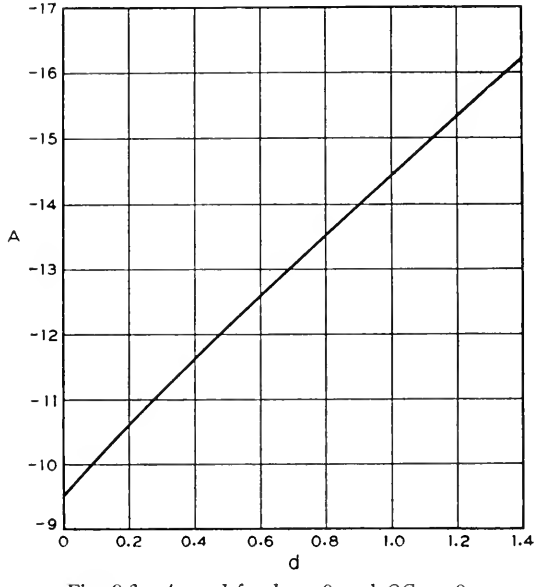

Fig. 9.3—A vs d for $b = 0$ and $QC = 0$.

In Fig. 9.2, A is plotted vs b for these three values of d.

It is perhaps of some interest to plot A vs d for $b = 0$ (the electron velocity equal to the phase velocity of the undisturbed wave). Such a plot is shown in Fig. 9.3.

9.4. Space Charge

We will now consider the case in which $QC \neq 0$. We will deal with this case only for $d = 0$, and for b adjusted for maximum gain per wavelength.

There is a peculiarity about this case in that a certain voltage V is applied to the *circuit* at $z = 0$, and we want to evaluate the *circuit* voltage associated with the increasing wave, V_{c1}, in order to know the gain.

At $z = 0$, $i = 0$. Now, the term which multiplies i to give the space-charge component of voltage (the second term on the right in (7.11)) is the same for all three waves and hence at $z = 0$ the circuit voltage is the total voltage. Thus, (9.1)–(9.3) hold. However, after V_1 has been obtained from (9.4), with

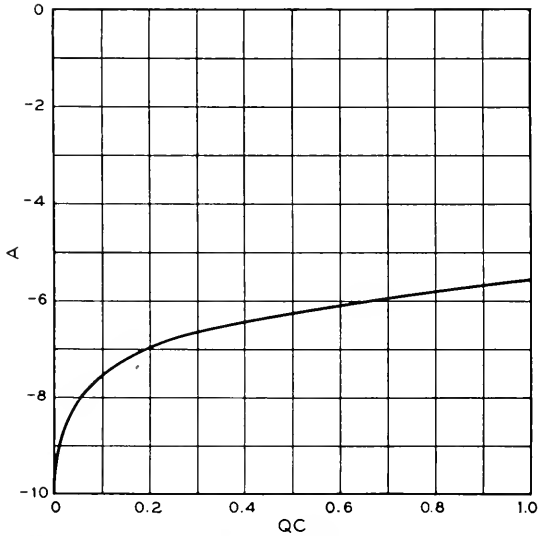

Fig. 9.4—A vs QC for $d = 0$ and b chosen for maximum gain of the increasing wave.

$V = V_1$, $v = i = 0$, then the circuit voltage V_{c1} must be obtained through the use of (7.14), and the initial loss parameter is

$$A = 20 \log_{10} | V_{c1}/V | \tag{9.15}$$

By using the appropriate values of δ, the same used in plotting Figs. 8.1 and 8.7–8.9, the loss parameter A was obtained from (9.15) and plotted vs QC in Fig. 9.4.

9.5 Change in Loss

We might think it undesirable in introducing loss to make the whole length of the helix lossy. For instance, we might expect the power output to be higher if the last part of the helix had low loss. Also, from Figs. 8.2

and 8.3 we see that the initial loss A becomes higher as d is increased. This is natural, because the electron stream can act to cause gain only after it is bunched, and if the initial section of the circuit is lossy, the signal decays before the stream becomes strongly bunched.

Let us consider a section of a lossless helix which is far enough from the input so that the increasing wave predominates and the total voltage V can be taken as that corresponding to the increasing wave

$$V = V_1 \tag{9.16}$$

Then, at this point

$$(ju_0C/\eta)v = V_1/\delta_1 \tag{9.17}$$

$$(-2V_0C^2/I_0)i = V_1/\delta_1^2 . \tag{9.18}$$

Here δ_1 is the value for $d = 0$ (and, we assume, $b = 0$). If we substitute the values from (9.16) in (9.4), and use in (9.4) the values of δ corresponding to $b = Q = 0$, $d \neq 0$, and call the value of V_1 we obtain V_1', we obtain the ratio of the initial amplitude of the increasing wave in the lossy section to the value of the increasing wave just to the left of the lossy section. Thus, the loss in the amplitude of the increasing wave in going from a lossless to a lossy section is $20 \log_{10} | V_1'/V_1 |$. This loss is plotted vs d in Fig. 8.5.

This loss is accounted for by the fact that $| i_1/V_1 |$ becomes larger as the loss parameter d is increased. Thus, the convection current injected into the lossy section is insufficient to go with the voltage, and the voltage must fall.

If we go from a lossy section ($d \neq 0$, $b = 0$) to a lossless section ($d = 0$, $b = 0$) we start with an excess of convection current and $| V_1' |$, the initial amplitude of the increasing wave to the right of the discontinuity is greater than the amplitude $| V_1 |$ of the increasing wave to the left. In Fig. 9.5, $20 \log_{10} | V_1'/V_1 |$ is plotted vs d for this case also.

We see that if we go from a lossless section to a lossy section, and if the lossy section is long enough so that the increasing wave predominates at the end of it, and if we go back to a lossless section at the end of it, the net loss and gain at the discontinuities almost compensate, and even for $d = 3$ the net discontinuity loss is less than 1 db. This does not consider the reduction of gain of the increasing wave in the lossy section.

9.6 Severed Helix

If the loss introduced is distributed over the length of the helix, the gain will decrease as the loss is increased (Fig. 8.5). If, however, the loss is distributed over a very short section, we easily see that as the loss is increased more and more, the gain must approach a constant value. The circuit will

be in effect severed as far as the electromagnetic wave is concerned, and any excitation in the output will be due to the a-c velocity and convection current of the electron stream which crosses the lossy section.

We will first idealize the situation and assume that the helix is severed and by some means terminated looking in each direction, so that the voltage falls from a value V to a value 0 in zero distance, while v and i remain unchanged.

We will consider a case in which $b = d = Q = 0$, and in which a voltage

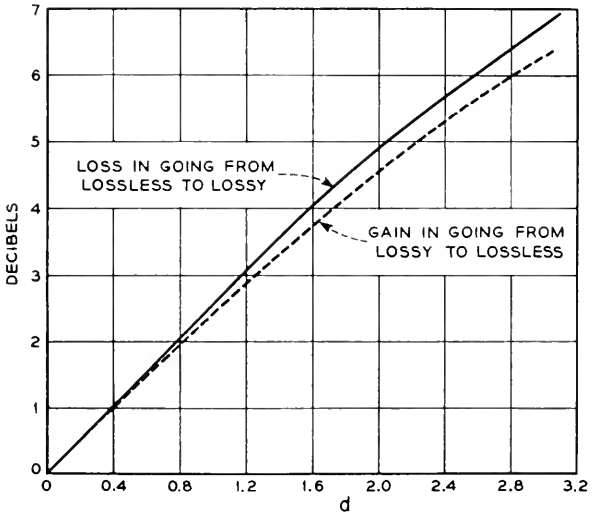

Fig. 9.5—Suppose that the circuit loss parameter changes suddenly with distance from 0 to d or from d to 0. Suppose there is an increasing wave only incident at the point of change. How large will the increasing wave beyond the point of change be? These curves tell ($b = QC = 0$).

V is applied to the helix N wavelengths before the cut. Then, just before the cut,

$$\begin{aligned} V_1 &= (V/3)e^{-j2\pi N} e^{2\pi NC\delta_1} \\ V_2 &= (V/3)e^{-j2\pi N} e^{2\pi NC\delta_2} \\ V_3 &= (V/3)e^{-j2\pi N} e^{2\pi NC\delta_3} \end{aligned} \tag{9.19}$$

and

$$\begin{aligned} (j\mu_0 C/\eta)v_1 &= V_1/\delta_1 \\ (-2V_0C^3/I_0)i_1 &= V_1/\delta_1^2 \end{aligned} \tag{9.20}$$

etc.

Whence, just beyond the break which makes $V = 0$, V, v and i are

$$V = 0$$

$$(ju_0C/\eta)v = V_1/\delta_1 + V_2/\delta_2 + V_3/\delta_3 \tag{9.21}$$

$$(-2V_0C^3/I_0)i = V_1/\delta_1^2 + V_2/\delta_2^2 + V_3/\delta_3^2$$

Putting these values in (9.4), we can find V_1', the value of the increasing wave to the right of the break. The ratio of the magnitude of the increasing wave to the magnitude it would have if it were not for the break is then $| V_1'/V_1 |$, and this ratio is plotted vs CN in Fig. 9.6, where N is the number of wavelengths in the first section.

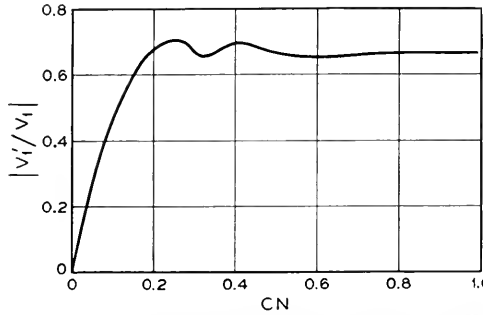

Fig. 9.6—Suppose the circuit is severed a distance measured by CN beyond the input, so that the voltage just beyond the break is zero. The ordinate is the ratio of the amplitude of the increasing wave beyond the break to that it would have had with an unbroken circuit ($b = QC = 0$).

We see that there will be least loss in severing the helix for CN equal to approximately $\frac{1}{4}$. From Fig. 9.1, we see that at $CN = \frac{1}{4}$ the voltage is just beginning to rise. In a typical 4,000 megacycle traveling-wave tube, CN is approximately unity for a 10 inch helix, so the loss should be put at least 2.5″ beyond the input. Putting the loss further on changes things little; asymptotically, $| V_1'/V |$ approaches $\frac{2}{3}$, or 3.52 db loss, for large values of CN (loss for from input).

It is of some interest to know how the voltage rises to the right of the cut. It was assumed that the cut was far from the point of excitation, so that only increasing wave of magnitude V_1 was present just to the left of the cut. The initial amplitudes of the three waves, V_1', V_2', V_3' to the right of the cut were computed and the magnitude of their sum plotted vs CN as it varies with distance to the right of the cut. The resulting curve, expressed in db with respect to the magnitude of the increasing wave V_1 just to the left of the cut, is shown in Fig. 9.7. Again, we see that at a distance $CN = \frac{1}{4}$ to the right of the cut the increasing wave (dashed straight line) predominates.

9.7 Severed Helix With Drift Space

In actually putting concentrated loss in a helix, the loss cannot be concentrated in a section of zero length for two reasons. In the first place, this is physically difficult if not impossible; in the second place it is desirable that the two halves of the helix be terminated in a reflectionless manner at the cut, and it is easiest to do this by tapering the loss. For instance, if the loss is put in by spraying aquadag (graphite in water) on ceramic rods supporting the helix, it is desirable to taper the loss coating at the ends of the lossy section.

Perhaps the best reasonably simple approximation we can make to such a lossy section is one in which the section starts far enough from the input

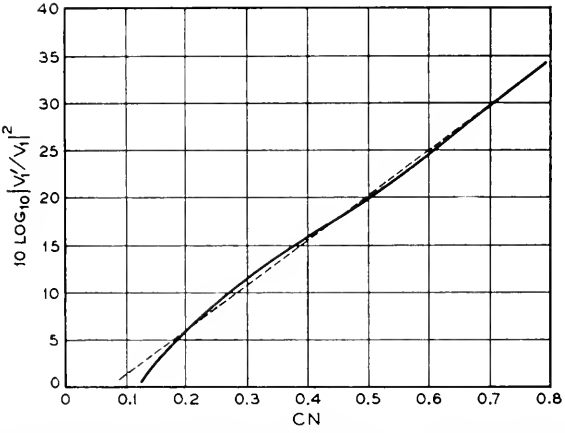

Fig. 9.7—Suppose that the circuit is severed and an increasing wave only is incident at the break. How does the signal build up beyond the break? The solid curve shows ($b = QC = 0$). 0 db is the level of the incident increasing wave.

so that at the beginning of the lossy section only an increasing wave is present. In the lossy section CN long we will consider that the loss completely shorts out the circuit, so that (8.28) holds. Thus, in the lossy section we will have only two values of δ, which we will call δ_I and δ_{II}.

$$\delta_I = jk \tag{9.21}$$

$$\delta_{II} = -jk \tag{9.22}$$

$$k = 2\sqrt{QC} \tag{9.23}$$

Let V_I and V_{II} be the voltages of the waves corresponding to δ_I and δ_{II} at the beginning of the lossy section. Let δ_1, δ_2, δ_3 be the values of δ to the left and right of the lossy section. Let V_1 be the amplitude of the increasing

wave just to the left of the lossy section. Then, by equating velocities and convection currents at the start of the lossy section, we obtain

$$V_1/\delta_1 = V_I/\delta_I + V_{II}/\delta_{II} \tag{9.24}$$

and, from (9.21) and (9.22)

$$V_1/\delta_1 = (-j/k)(V_I - V_{II}) \tag{9.25}$$

Similarly

$$\begin{aligned} V_1/\delta_1^2 &= V_I/\delta_{II}^2 + V_{II}\delta_{II}^2 \\ V_1\delta_1^2 &= -(1/k^2)(V_I + V_{II}) \end{aligned} \tag{9.26}$$

So that

$$V_I = j(V_1/2)(k/\delta_1)(jk/\delta_1 + 1) \tag{9.27}$$

$$V_{II} = j(V_1/2)(k/\delta_1)(jk/\delta_1 - 1) \tag{9.28}$$

At the output of the lossy section we have the voltages V_I' and V_{II}'

$$V_I' = V_I e^{-j2\pi N} e^{-j2\pi kCN} \tag{9.29}$$

$$V_{II}' = V_{II} e^{-j2\pi N} e^{-j2\pi kCN} \tag{9.30}$$

Thus, at the end of the lossy section we have

$$V = V_I' + V_{II}' \tag{9.31}$$

$$\begin{aligned} (ju_0C/\eta)v &= V_I'/\delta_I + V_{II}'/\delta_{II} \\ (ju_0C/\eta)v &= (-j/k)(V_I' - V_{II}') \end{aligned} \tag{9.32}$$

and similarly

$$(-2V_0C^2/I_0)i = (-1/k^2)(V_I' + V_{II}') \tag{9.33}$$

From (9.27) and (9.28) we see that

$$V_I' + V_{II}' = -(k/\delta_1)[+(k/\delta_1)\cos 2\pi kCN + \sin 2\pi kCN]V_1 e^{-j2\pi N} \tag{9.34}$$

$$V_I' - V_{II}' = j(k/\delta_1)[-(k/\delta_1)\sin 2\pi kCN + \cos 2\pi kCN]V_1 e^{-j2\pi N} \tag{9.35}$$

Whence

$$V = -(k/\delta_1)[+(k/\delta_1)\cos 2\pi kCN + \sin 2\pi kCN]V_1 e^{-j2\pi N} \tag{9.36}$$

$$(ju_0C/\eta)v = (1/\delta_1)[-(k/\delta_1)\sin 2\pi kCN + \cos 2\pi kCN]V_1 e^{-j2\pi N} \tag{9.37}$$

$$(-2V_0C^2/I_0)i = (1/\delta_1)[(1/\delta_1)\cos 2\pi kCN + (1/k)\sin 2\pi kCN]V_1 e^{-j2\pi N} \tag{9.38}$$

These can be used in connection with (9.4) in obtaining V_1', the value of V_1 just beyond the lossy section; that is, the amplitude of the component of increasing wave just beyond the lossy section.

In typical traveling-wave tubes the lossy section usually has a length such that CN is $\frac{1}{4}$ or less. In Fig. 9.8 the loss in db in going through the lossy section, $20 \log_{10} | V_1'/V_1 |$, has been plotted vs. CN for $QC = 0, .25, .5$ for the range $CN = 0$ to $CN = .5$.

We see that, for low space charge, increasing the length of a drift space increases the loss. For higher space charge it may either increase or decrease the loss. It is not clear that the periodic behavior characteristic of the curves for $QC = 0.5$ and 1, for instance, will obtain for a drift space with tapered loss at each end. The calculations may also be considerably in error for broad electron beams (γa large). The electric field pattern in the helix differs

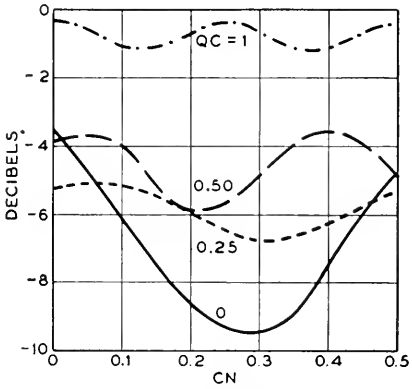

Fig. 9.8—Suppose that we break the circuit and insert a drift tube of length measured by CN in terms of the traveling-wave tube C and N. Assume an increasing wave only before the drift tube. The increasing wave beyond the drift tube will have a level with respect to the incident increasing wave as shown by the ordinate. Here $d = 0$ and b is chosen to maximize x_1.

from that in the drift space. In the case of broad electron beams this may result in the excitation in the drift space of several different space charge waves having different field patterns and different propagation constants.

A suggestion has been made that the introduction of loss itself has a bad effect. The only thing that affects the electrons is an electric field. Unpublished measurements made by Cutler mo le by moving a probe along a helix indicate that in typical short high-loss sections the electric field of the helix is essentially zero. Hence, except for a short distance at the ends, such lossy sections should act simply as drift spaces.

9.8 Overall Behavior of Tubes

The material of Chapters VIII and IX is useful in designing traveling-wave tubes. Prediction of the performance of a given tube over a wide range of voltage and current is quite a different matter. For instance, in order to predict gain for voltage or current ranges for which the gain is small, the

three waves must be taken into account. As current is varied, the loss parameter d varies, and this means different x's and y's must be computed for different currents. Finally, at high currents, the space-charge parameter Q must be taken into account. In all, a computation of tube behavior under a variety of conditions is an extensive job.

Fortunately, for useful tubes operating as intended, the gain is high. When this is so, the gain can be calculated quite accurately by asymptotic relations. Such an overall calculation of the gain of a helix-type tube with distributed loss is summarized in Appendix VII.

CHAPTER X

NOISE FIGURE

Synopsis of Chapter

BECAUSE THERE IS no treatment of the behavior at high frequencies of an electron flow with a Maxwellian distribution of velocities, one might think there could be no very satisfactory calculation of the noise figure of traveling-wave tubes. Various approximate calculations can be made, and two of these will be discussed here. Experience indicates that the second and more elaborate of these is fairly well founded. In each case, an approximation is made in which the actual multi-velocity electron current is replaced by a current of electrons having a single velocity at a given point but having a mean square fluctuation of velocity or current equal to a mean square fluctuation characteristic of the multi-velocity flow.

In one sort of calculation, it is assumed that the noise is due to a current fluctuation equal to that of shot noise (equation (10.1)) in the current entering the circuit. For zero loss, an electron velocity equal to the phase velocity of the circuit and no space charge, this leads to an expression for noise figure (10.5), which contains a term proportional to beam voltage V_0 times the gain parameter C. One can, if he wishes, add a space-charge noise reduction factor multiplying the term $80\,V_0C$. This approach indicates that the voltage and the gain per wavelength should be reduced in order to improve the noise figure.

In another approach, equations applying to single-valued-velocity flow between parallel planes are assumed to apply from the cathode to the circuit, and the fluctuations in the actual multi-velocity stream are represented by fluctuations in current and velocity at the cathode surface. It is found that for space-charge-limited emission the current fluctuation has no effect, and so all the noise can be expressed in terms of fluctuations in the velocity of emission of electrons.

For a special case, that of a gun with an anode at circuit potential V_0, a cathode-anode transit angle θ_1, and an anode-circuit transit angle θ_2, an expression for noise figure (10.28) is obtained. This expression can be rewritten in terms of a parameter L which is a function of P

$$F = 1 + (\tfrac{1}{2})(4 - \pi)(T_c/T)(1/C)L$$

$$P = (\theta_1 - \theta_2)C$$

Formally, F can be minimized by choosing the proper value of P. In Fig. 10.3, the minimum value of L, L_m, is plotted vs. the velocity parameter b for zero loss and zero space charge ($d = QC = 0$). The corresponding value of P, P_m, is also shown.

P is a function of the cathode-anode transit angle θ_1, which cannot be varied without changing the current density and hence C, and of anode-circuit transit angle θ_2, which can be given any value. Thus, P can be made very small if one wishes, but it cannot be made indefinitely large, and it is not clear that P can always be made equal to P_m. On the other hand, these expressions have been worked out for a rather limited case: an anode potential equal to circuit potential, and no a-c space charge. It is possible that an optimization with respect to gun anode potential and space charge parameter QC would predict even lower noise figures, and perhaps at attainable values of the parameters.

In an actual tube there are, of course, sources of noise which have been neglected. Experimental work indicates that partition noise is very important and must be taken into account.

10.1 Shot Noise in the Injected Current

A stream of electrons emitted from a temperature-limited cathode has a mean square fluctuation in convection current $\overline{i_s^2}$

$$\overline{i_s^2} = 2eI_0B_0 \tag{10.1}$$

Here e is the charge on an electron, I_0 is the average or d-c current and B is the bandwidth in which the frequencies of the current components whose mean square value is $\overline{i_s^2}$ lie. Suppose this fluctuation in the beam current of a traveling-wave tube were the sole cause of an increasing wave ($V = v = 0$). Then, from (9.4) the mean square value of that increasing wave,, $\overline{V_{1s}^2}$, would be

$$\overline{V_{1s}^2} = (8eBV_0^2C^4/I_0) \mid \delta_2\delta_3 \mid^2 \mid (1 - \delta_2/\delta_1)(1 - \delta_3/\delta_1) \mid^{-2} \tag{10.2}$$

Now, suppose we have an additional noise source: thermal noise voltage applied to the circuit. If the helix is matched to a source of temperature T, the thermal noise power P_t drawn from the source is

$$P_t = kTB \tag{10.3}$$

Here k is Boltzman's constant, T is temperature in degrees Kelvin and, as before, B is bandwidth in cycles. If K_ℓ is the longitudinal impedance of the circuit the mean square noise voltage $\overline{V_t^2}$ associated with the circuit will be

$$\overline{V_t^2} = kTBK_\ell \tag{10.4}$$

and the component of increasing wave excited by this voltage, $\overline{V_{1t}^2}$, will be, from (9.4),

$$\overline{V_{1t}^2} = kTBK_t \mid (1 - \delta_2/\delta_1)(1 - \delta_3/\delta_1) \mid^{-2} \tag{10.5}$$

The noise figure of an amplifier is defined as the ratio of the total noise output power to the noise output power attributable to thermal noise at the input alone. We will regard the mean-square value of the initial voltage V_1 of the increasing wave as a measure of noise output. This will be substantially true if the signal becomes large prior to the introduction of further noise. For example, it will be substantially true in a tube with a severed helix if the helix is cut at a point where the increasing wave has grown large compared with the original fluctuations in the electron stream which set it up.

Under these circumstances, the noise figure F will be given by

$$F = \overline{(V_{1s}^2 + V_{1t}^2)}/\overline{(V_{1t}^2)}$$

$$F = 1 + (e/kT)(8V_0^2C^4/I_0K_t) \mid \delta_2\delta_3 \mid^2 \tag{10.3}$$

Now we have from Chapter II that

$$C^3 = I_0K_t/4V_0$$

whence

$$F = 1 + 2(eV_0/kT)C \mid \delta_2\delta_3 \mid^2 \tag{10.4}$$

The standard reference temperature is 290°K. Let us assume $b = d = QC = 0$. For this case we have found $\mid \delta_2 \mid = \mid \delta_3 \mid = 1$. Thus, for these assumptions we find

$$F = 1 + 80V_0C \tag{10.5}$$

A typical value of V_0 is 1,600 volts; a typical value of C is .025. For these values

$$F = 3{,}201$$

In db this is a noise of 35 db.

This is not far from the noise figure of traveling-wave tubes when the cathode temperature is lowered so as to give temperature-limited emission. The noise figure of traveling-wave tubes in which the cathode is at normal operating temperature and is active, so that emission is limited by space-charge, can be considerably lower. In endeavoring to calculate the noise figure for space-charge-limited electron flow from the cathode we must proceed in a somewhat different manner.

10.2 The Diode Equations

Llewellyn and Peterson[1] have published a set of equations governing the behavior of parallel plane diodes with a single-valued electron velocity. They sum up the behavior of such a diode in terms of nine coefficients A^*–I^*, in the following equations

$$V_b - V_a = A^* I + B^* q_a + C^* v_a \tag{10.6}$$

$$q_b = D^* I + E^* q_a + F^* v_a \tag{10.7}$$

$$v_b = G^* I + H^* q_a + I^* v_a \tag{10.8}$$

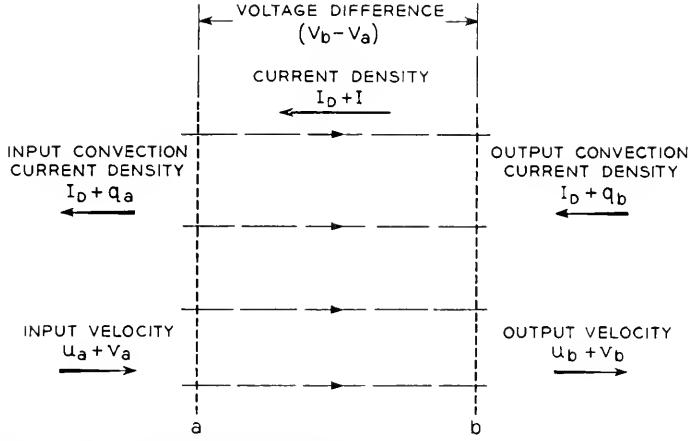

Fig. 10.1—Parallel electron flow between two planes a and b normal to the flow, showing the currents, velocities and voltages.

These equations and the values of the various coefficients in terms of current, electron velocity and transit angle are given in Appendix V. The diode structure to which they apply is indicated in Fig. 10.1. Electrons enter normal to the left plane and pass out at the right plane. The various quantities involved are transit angle between the two planes and:

I_0	d-c current density to left
I	a-c current density to left
q_a	a-c convection current density to left at input plane a
q_b	a-c convection current density to left at output plane b
u_a	d-c velocity to right at plane a
u_b	d-c velocity to right at plane b
v_a	a-c velocity to right at plane a
v_b	a-c velocity to right at plane b
V_b-V_a	a-c potential difference between plane b and plane a

[1] F. B. Llewellyn and L. C. Peterson, "Vacuum Tube Networks," *Proc. I.R.E.*, Vol. 32, pp. 144–166, March, 1944.

We will notice that I and the q's are current *densities* and that, contrary to the convention we have used, they are taken as positive to the left. Thus, if the area is σ, we would write the output convection current; as

$$i = -\sigma q_b$$

where q_b is the convection current density used in (10.6)–(10.8).

Peterson has used (10.6)–(10.8) in calculating noise figure by replacing the actual multi-velocity flow from the cathode by a single-velocity flow with the same mean square fluctuation in velocity, namely,[2]

$$\overline{v_t^2} = (4 - \pi)\eta\,(kT_c/I_0)B \tag{10.9}$$

Here T_c is the cathode temperature in degrees Kelvin and I_0 is the cathode current.

Whatever the justification for such a procedure, Rack[3] has shown that it gives a satisfactory result at low frequencies, and unpublished work by Cutler and Quate indicates surprisingly good quantitative agreement under conditions of long transit angle at 4,000 mc.

We must remember, however, that the available values of the coefficients of (10.6)–(10.8) are for a broad electron beam in which there are a-c fields in the z direction only. Now, the electron beam in the gun of a traveling-wave tube is ordinarily rather narrow. While the a-c fields may be substantially in the z-direction near the cathode, this is certainly not true throughout the whole cathode-anode space. Thus, the coefficients used in (10.6)–(10.8) are certainly somewhat in error when applied to traveling-wave tube guns.

Various plausible efforts can be made to amend this situation, as, by saying that the latter part of the beam in the gun acts as a drift region in which the electron velocities are not changed by space-charge fields. However, when one starts such patching, he does not know where to stop. In the light of available knowledge, it seems best to use the coefficients as they stand for the cathode-anode region of the gun.

Let us then consider the electron gun of the traveling-wave tube to form a space-charge limited diode which is short-circuited at high frequencies.

If we assume complete space charge (space-charge limited emission) and take the electron velocity at the cathode to be zero, we find that the quantities multiplying q_a in (10.6)–(10.8) are zero.

$$B^* = E^* = H^* = 0^* \tag{10.10}$$

[2] L. C. Peterson, "Space-Charge and Transit-Time Effects on Signal and Noise in Microwave Tetrodes," *Proc. I.R.E.*, Vol. 35, pp. 1264–1272, November, 1947.

[3] A. J. Rack, "Effect of Space Charge and Transit Time on the Shot Noise in Diodes," *Bell System Technical Journal*, Vol. 17, pp. 592–619, October, 1938.

Accordingly, the magnitude of the noise convection current at the cathode does not matter. If we assume that the gun is a short-circuited diode as far as r-f goes

$$V_b - V_a = 0 \tag{10.11}$$

Then from (10.6), (10.10) and (10.11) we obtain

$$I = -\frac{C^*}{A^*} v_a \tag{10.12}$$

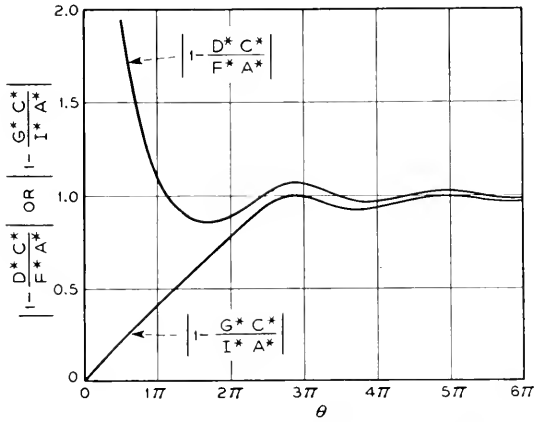

Fig. 10.2—Some expressions useful in noise calculations, showing how they approach unity at large transit angles.

Accordingly, from (10.7) and (10.8) we obtain

$$q_b = \left(1 - \frac{D^*C^*}{F^*A^*}\right) F^* v_a \tag{10.13}$$

$$v_b = \left(1 - \frac{G^*C^*}{I^*A^*}\right) I^* v_a \tag{10.14}$$

In Fig. 10.2, $|1 - D^*C^* F^*A^*|$ and $|1 - G^*C^* I^*A^*|$ are plotted vs θ, the transit angle. We see that for transit angles greater than about 3π these quantities differ negligibly from unity, and we may write

$$q_b = F^* v_a \tag{10.15}$$

$$v_b = I^* v_a \tag{10.16}$$

More specifically, we find

$$q_b = \frac{v_a I_0 \beta_1 e^{-\beta_1}}{u_b} \tag{10.17}$$

$$v_b = -v_a e^{-\beta_1} \tag{10.18}$$

Here β_1 is j times the transit angle in radians from cathode to anode. For v_a we use a velocity fluctuation with the mean-square value given by (10.9).

Suppose now that there is a constant-potential drift space following the diode anode, of length β_2/j in radians. If we apply (10.6)–(10.8) and assume that the space-charge is small and the transit angle long, we find that q_b', the value of q_b at the end of this drift space, is given in terms of q_a' and v_a', the values at the beginning of this drift space, by

$$q_b' = (q_a' + (I_0/u_b)\beta_2 v_a')e^{-\beta_2} \tag{10.19}$$

The case of v_b', the velocity at the end of this drift space, is a little different. The first term on the right of (10.8) can be shown to be negligible for long transit angles and small space charge. The last term on the right represents the purely kinematic bunching. For the assumption of small space charge the middle term gives not zero but a first approximation of a space-charge effect, assuming that all the space-charge field acts longitudinally. Thus, this middle term gives an overestimate of the effect of space-charge in a narrow, high-velocity beam. If we include both terms, we obtain

$$v_b' = H_2^* q_a' + e^{-\beta_2} v_a' \tag{10.20}$$

Here the term on the right is the purely kinematic term.*

Now, the current from the gun is assumed to go into the drift space, so that q_a' is q_b from (10.17) and v_a' is v_a from (10.18). The d-c velocity at the gun anode and throughout the drift space are both given by u_b. If we make these substitutions in (10.19) and (10.20) we obtain

$$q_b' = (I_0/u_b)(\beta_1 - \beta_2)e^{-(\beta_1+\beta_2)} v_a \tag{10.21}$$

$$v_b' = -\left(2\frac{\beta_1}{\beta_2} + 1\right) e^{-(\beta_1+\beta_2)} v_a \tag{10.22}$$

The term $2\beta_1/\beta_2$ in (10.22) is the "space-charge" term. We will in the following analysis omit this, making the same sort of error we do in neglecting space charge in the traveling-wave section of the tube. If space charge in the drift space is to be taken into account, it is much better to proceed as in 9.7.

From the drift-space the current goes into the helix. It is now necessary to change to the notation we have used in connection with the traveling-wave tube. The chief difference is that we have taken currents as positive to the right, but allowed I_0 to be the d-c current to the left. If i and v are

* The first term has been written as shown because it is easiest to use the small space-charge value of H^* for the drift region (H_2^*) in connection with the space-charge limited value of F^* for the cathode-anode region rather than in connection with (10.17).

our a-c convection current and velocity at the beginning of the helix, and I_0 and u_0 the d-c beam current and velocity, and σ the area of the beam,

$$\begin{aligned} i &= -\sigma q_b' \\ v &= v_b \\ I_0 &= \sigma I_0 \\ u_0 &= u_0 \end{aligned} \tag{10.23}$$

In addition, we will use transit angles θ_1 and θ_2 in place of β_1 and β_2

$$\begin{aligned} \beta_1 &= j\theta_1 \\ \beta_2 &= j\theta_2 \end{aligned} \tag{10.24}$$

We then obtain from (10.21) and (10.22)

$$q = -j(I_0/u_0)(\theta_1 - \theta_2)e^{-j(\theta_1+\theta_2)}v_a \tag{10.25}$$

$$v = -e^{-j(\theta_1+\theta_2)}v_a \tag{10.26}$$

10.3 Overall Noise Figure

We are now in a position to use (9.4) in obtaining the overall noise figure. We have already assumed that the space-charge is small in the drift space between the gun anode and the helix ($QC = 0$). If we continue to assume this in connection with (9.4), the only voltage is the helix voltage and for the noise caused by the velocity fluctuation at the cathode, v_a, $V = 0$ at the beginning of the helix. Thus, the mean square initial noise voltage of the increasing wave, $\overline{V_{1s}^2}$, will be, from (10.21), (10.22), (9.4) and (10.9),

$$\overline{V_{1s}^2} = (2(4 - \pi)kT_eCBV_0/I_0)|\,\delta_2\delta_3(\theta_1 - \theta_2)C + (\delta_2 + \delta_3)\,|^2 \\ |\,(1 - \delta_2\delta_1)(1 - \delta_3\delta_1)\,|^{-2} \tag{10.27}$$

As before, we have, from the thermal noise input to the helix

$$\overline{V_{1t}^2} = kTBK_t|\,(1 - \delta_2\delta_1)(1 - \delta_3\delta_1)\,|^{-2} \tag{10.5}$$

and the noise figure becomes

$$F = 1 + \overline{V_{1s}^2}/\overline{V_{1t}^2}$$

$$F = 1 + (1/2)(4 - \pi)(T_e/T)(1/C)|\,\delta_2\delta_3(\theta_1 - \theta_2)C + (\delta_2 + \delta_3)\,|^2 \tag{10.28}$$

Here use has been made of the fact that

$$C = K_tI/4V_0$$

Let us investigate this for the case $b = d = 0$ (we have already assumed $QC = 0$). In this case

$$\delta_2 = \sqrt{3}/2 - j1/2$$

$$\delta_3 = j$$

and we obtain

$$F = 1 + (1/2)(4 - \pi)(T_c/T)(1/C)|(P/2 - \sqrt{3}/2) - j(\sqrt{3}P/2 - 1/2)|^2 \tag{10.29}$$

$$P = (\theta_1 - \theta_2)C \tag{10.30}$$

For a given gun transit-angle θ_1, the parameter P can be given values ranging from $\theta_1 C$ to large negative values by increasing the drift angle θ_2 between the gun anode and the beginning of the helix.

We see that

$$F = 1 + (1/2)(4 - \pi)(T_c/T)(1/C)(P^2 - \sqrt{3}P + 1) \tag{10.31}$$

The minimum value of $(P^2 - \sqrt{3}P + 1)$ occurs when

$$P = \sqrt{3}/2 \tag{10.32}$$

if the product of the gun transit angle and C is large enough, this can be attained. The corresponding value of $(P^2 - \sqrt{3}P + 1)$ is $\frac{1}{4}$, and the corresponding noise figure is

$$F = 1 + (1/2)(1 - \pi/4)(T_c/T)(1/C) \tag{10.33}$$

A typical value for T_c is $1020°K$, and for a reference temperature of $290°K$,

$$T_c/T = 3.5$$

A typical value of C is .025. For these values

$$F = 17$$

or a noise figure of 12 db.

Let us consider cases for no attenuation or space-charge but for other electron velocities. In this case we write, as before

$$\delta_2 = x_2 + jy_2$$

$$\delta_3 = x_3 + jy_3$$

Let us write, for convenience,

$$L = |\delta_2\delta_3 P + \delta_1 + \delta_2|^2 \tag{10.34}$$

Then we find that

$$L = [(x_2x_3)^2 + (y_2y_3)^2 + (x_2y_3)^2 + (x_3y_2)^2]P^2 + 2[x_3(y_2^2 + x_2^2) + x_2(x_3^2 + y_3^2)]P + (x_2 + x_3)^2 + (y_2 + y_3)^2 \tag{10.35}$$

This has a minimum value for $P = P_m$

$$P_m = \frac{-[x_3(x_2^2 + y_2^2) + x_2(x_3^2 + y_3^2)]}{(x_2x_3)^2 + (y_2y_3)^2 + (x_2y_3)^2 + (x_3y_2)^2} \tag{10.36}$$

We note that, as we are not dealing with the increasing wave, x_2 and x_3

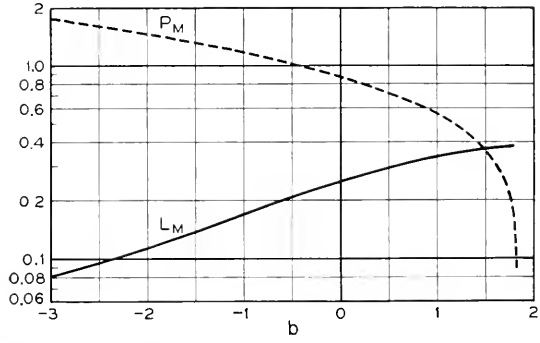

Fig. 10.3—According to the theory presented, the overall noise figure of a tube with a lossless helix and no space charge is proportional to L. Here we have a minimum value of L_m, minimized with respect to P, which is dependent on gun transit angle, and also the corresponding value of P, P_m. According to this curve, the optimum noise figure should be lowest for low electron velocities (low values of b). It may, however, be impossible to make P equal to P_m.

must be either negative or zero, and hence P_m is always positive. For no space-charge and no attenuation, x_3 is zero for all values of b and

$$P_m = \frac{-x_2}{y_2^2 + x_2^2} \tag{10.37}$$

From (10.36) and (10.35), the minimum value of L, L_m, is

$$L_m = (x_2 + x_3)^2 + (y_2 + y_3)^2 - \frac{[x_3(y_2^2 + x_2^2) + x_2(x_3^2 + y_3^2)]^2}{(x_2x_3)^2 + (y_2y_3)^2 + (x_2y_3)^2 + (x_3y_2)^2} \tag{10.38}$$

When $x_3 = 0$, as in (10.37)

$$L_m = x_2^2 + y_2^2 + 2y_2y_3 + \frac{y_2^2y_3^2}{x_2^2 + y_2^2} \tag{10.39}$$

In Fig. 10.3, P_m and L_m are plotted vs b for no attenuation ($d = 0$). We see that P_m becomes very small as b approaches $(3/2)2^{1/3}$, the value at which the increasing wave disappears.

If space charge is to be taken into account, it should be taken into account both in the drift space between anode and helix and in the helix itself. In the helix we can express the effect of space-charge by means of the parameter QC and boundary conditions can be fitted as in Chapter IX. The drift space can be dealt with as in Section 9.7 of Chapter IX. The inclusion of the effect of space-charge by this means will of course considerably complicate the analysis, especially if $b \neq 0$.

While working with Field at Stanford, Dr. C. F. Quate extended the theory presented here to include the effect of all three waves in the case of low gain, and to include the effect of a fractional component of beam current having pure shot noise, which might arise through failure of space-charge reduction of noise toward the edge of the cathode. His extended theory agreed to an encouraging extent with his experimental results. Subsequent unpublished work carried out at these Laboratories by Cutler and Quate indicates a surprisingly good agreement between calculations of this sort and observed noise current, and emphasizes the importance of properly including both partition noise and space charge in predicting noise figure.

10.4 Other Noise Considerations

Space-charge reduction of noise is a cooperative phenomenon of the whole electron beam. If some electrons are eliminated, as by a grid, additional "partition" noise is introduced. Peterson shows how to take this into account.[2]

An electron may be ineffective in a traveling-wave tube not only by being lost but by entering the circuit near the axis where the r-f field is weak rather than near the edge where the r-f field is high. Partition noise arises because sidewise components of thermal velocity cause a fluctuation in the amount of current striking a grid or other intercepting circuit. If such sidewise components of velocity appreciably alter electron position in the helix, a noise analogous to partition noise may arise even if no electrons actually strike the helix. Such a noise will also occur if the "counteracting pulses" of low-charge density which are assumed to smooth out the electron flow are broad transverse to the beam.

These considerations lead to some maxims in connection with low-noise traveling-wave tubes: (1) do not allow electrons to be intercepted by various electrodes (2) if practical, make sure that $I_0(\beta r)$ is reasonably constant over the beam, and/or (3) provide a very strong magnetic focusing field, so that electrons cannot move appreciably transversely.

10.5 Noise in Transverse-Field Tubes

Traveling-wave tubes can be made in which there is no longitudinal field component at the nominal beam position. One can argue that, if a narrow, well-collimated beam is used in such a tube, the noise current in the beam can induce little noise signal in the circuit (none at all for a beam of zero thickness with no sidewise motion). Thus, the idea of using a transverse-field tube as a low-noise tube is attractive. So far, no experimental results on such tubes have been announced.

A brief analysis of transverse-field tubes is given in Chapter XIII.

CHAPTER XI

BACKWARD WAVES

WE NOTED IN CHAPTER IV that, in filter-type circuits, there is an infinite number of spatial harmonics which travel in both directions. Usually, in a tube which is designed to make use of a given forward component the velocity of other forward components is enough different from that of the component chosen to avoid any appreciable interaction with the electron stream. It may well be, however, that a backward-traveling component has almost the same speed as a forward-traveling component.

Suppose, for instance, that a tube is designed to make use of a given forward-traveling component of a forward wave. Suppose that there is a forward-traveling component of a backward wave, and this forward-traveling component is also near synchronism with the electrons. Does this mean that under these circumstances both the backward-traveling and the forward-traveling waves will be amplified?

The question is essentially that of the interaction of an electron stream with a circuit in which the phase velocity is in step with the electrons but the group velocity and the energy flow are in a direction contrary to that of electron motion.

We can most easily evaluate such a situation by considering a distributed circuit for which this is true. Such a circuit is shown in Fig. 11.1. Here the series reactance X per unit length is negative as compared with the more usual circuit of Fig. 11.2. In the circuit of Fig. 11.2, the phase shift is 0° per section at zero frequency and assumes positive values as the frequency is increased. In the circuit of Fig. 11.1 the phase shift is −180° per section at a lower cutoff frequency and approaches 0° per section as the frequency approaches infinity.

Suppose we consider the equations of Chapter II. In (2.9) we chose the sign of X in such a manner as to make the series reactance positive, as in Fig. 11.2, rather than negative, as in Fig. 11.1. All the other equations apply equally well to either circuit. Thus, for the circuit of Fig. 11.1, we have, instead of (2.10),

$$V = \frac{+\Gamma\Gamma_1 i}{(\Gamma^2 - \Gamma_1^2)} \tag{11.1}$$

The sign is changed in the circuit equation relating the convection current and the voltage. Similarly, we can modify the equations of Chapter VII,

(7.9) and (7.12), by changing the sign of the left-hand side. From Chapter VIII, the equation for a lossless circuit with no space charge is

$$\delta^2(\delta + jb) = -j \tag{8.1}$$

The corresponding modification is to change the sign preceding δ^2, giving

$$\delta^2(\delta + jb) = +j \tag{11.2}$$

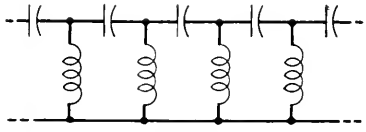

Fig. 11.1

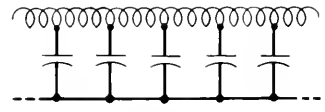

Fig. 11.2

Fig. 11.1—A circuit with a negative phase velocity. The electrons can be in synchronism with the field only if they travel in a direction opposite to that of electromagnetic energy flow.

Fig. 11.2—A circuit with a positive phase velocity.

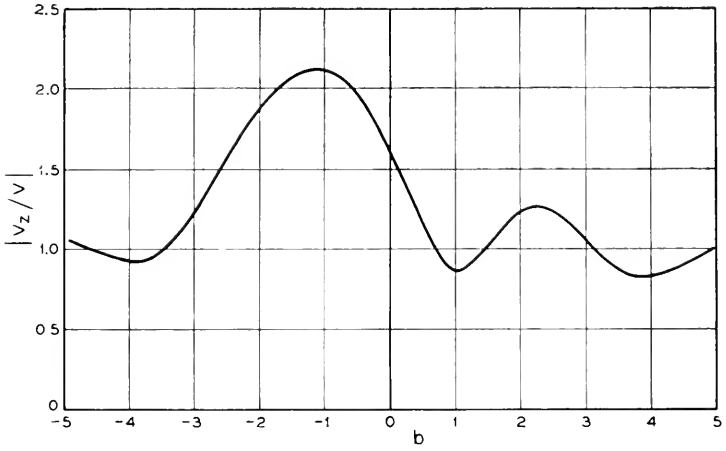

Fig. 11.3—Suppose we have a tube with a circuit such as that of Fig. 11.1, in which the circuit energy is really flowing in the opposite direction from the electron motion. Here, for $QC = d = 0$, we have the ratio of the magnitude of the voltage V_z a distance z from the point of injection of electrons to the magnitude of the voltage V at the point of injection of electrons. V_z is really the input voltage, and there will be gain at values of b for which $|V_z/V| < 1$.

In (11.2), b and δ have the usual meaning in terms of electron velocity and propagation constant.

Now consider the equation

$$\delta^2(\delta - jk) = j \tag{11.3}$$

Equations (11.2) and (8.1) apply to different systems. We have solutions of (8.1) and we want solutions of (11.2). We see that a solution of (11.2)

is a solution of (11.3) for $k = -b$. We see that a solution of (11.3) is the conjugate of a solution of (8.1) if we put b in (8.1) equal to k in (11.3). Thus, a solution of (11.2) is the conjugate of a solution of (8.1) in which b in (8.1) is made the negative of the value of b for which it is desired to solve (11.2).

We can use the solutions of Fig. 8.1 in connection with the circuit of Fig. 11.1 in the following way: wherever in Fig. 8.1 we see b, we write in instead $-b$, and wherever we see y_1, y_2 or y_3 we write in instead $-y_1$, $-y_2$ or $-y_3$.

Thus, for synchronous velocity, we have

$$\delta_1 = \sqrt{3}/2 + j1/2$$

$$\delta_2 = -\sqrt{3}/2 + j1/2$$

$$\delta_3 = -j$$

We can determine what will happen in a physical case only by fitting boundary conditions so that at $z = 0$ the electron stream, as it must, enters unmodulated.

Let us, for convenience, write Φ for the quantity βCz

$$\beta Cz = \Phi \tag{11.4}$$

We will have for the total voltage V_z at z in terms of the voltage V at $z = 0$

$$\begin{aligned} V_z = Ve^{-j\beta z}([(1 - \delta_2/\delta_1)(1 - \delta_3/\delta_1)]^{-1}e^{-j\Phi y_1}e^{\Phi x_1} \\ + [(1 - \delta_3/\delta_2)(1 - \delta_1/\delta_2)]^{-1}e^{-j\Phi y_2}e^{\Phi x_2} \\ + [(1 - \delta_1/\delta_3)(1 - \delta_2/\delta_3)]^{-1}e^{-j\Phi y_3}e^{\Phi x_3}) \end{aligned} \tag{11.5}$$

We must remember that in using values from an unaltered Fig. 8.1 we use in the δ's and as the y's the negative of the y's shown in the figure (the sign of the x's is unchanged), and for a given value of b we enter Fig. 8.1 at $-b$.

In Fig. 11.3, $|V_z/V|$ has been plotted vs b for $\Phi = 2$. We see that, for several values of b, $|V_z|$ (the input voltage) is less than $|V|$ (the output voltage) and hence there can be "backward" gain.

We note that as Φ is made very large, the wave which increases with increasing Φ will eventually predominate, and $|V_z|$ will be greater than $|V|$. "Backward gain" occurs not through a "growing wave" but rather through a sort of interference between wave components, as exhibited in Fig. 11.3.

Fig. 11.3 is for a lossless circuit; the presence of circuit attenuation would alter the situation somewhat.

APPENDIX IV

EVALUATION OF SPACE—CHARGE PARAMETER Q

Consider the system consisting of a conducting cylinder of radius a and an internal cylinder of current of radius a_1 with a current

$$ie^{j\omega t}e^{-\gamma z}. \tag{1}$$

Let subscript 1 refer to inside and 2 to outside. We will assume magnetic fields of the form

$$H_{\varphi 1} = AI_1(\gamma r) \tag{2}$$

$$H_{\varphi 2} = BI_1(\gamma r) + CK_1(\gamma r) \tag{3}$$

From Maxwell's equations we have,

$$\frac{\partial}{\partial r}(rH_\varphi) = j\omega\epsilon rE_z + rJ_z \tag{4}$$

Now

$$\frac{\partial}{\partial z}(zI_1(z)) = zI_0(z) \tag{5}$$

$$\frac{\partial}{\partial z}(zK_1(z)) = -zK_0(z) \tag{6}$$

Hence

$$E_{z1} = \frac{-j\gamma}{\omega\epsilon} AI_0(\gamma r) \tag{7}$$

$$E_{z2} = \frac{-j\gamma}{\omega\epsilon}(BI_0(\gamma r) - CK_0(\gamma r)) \tag{8}$$

at $r = a$, $E_{z2} = 0$

$$C = B\frac{I_0(\gamma a)}{K_0(\gamma a)} \tag{9}$$

at $r = a_1$, $E_{z1} = E_{z2}$

$$\begin{aligned} AI_0(\gamma a_1) &= B\left(I_0(\gamma a_1) - \frac{I_0(\gamma a)}{K_0(\gamma a)} K_0(\gamma a_1)\right) \\ A &= B\left(1 - \frac{I_0(\gamma a)}{K_0(\gamma a)} \frac{K_0(\gamma a_1)}{I_0(\gamma a_1)}\right) \end{aligned} \tag{10}$$

In going across boundary, we integrate (4) over the infinitesimal radial distance which the current is assumed to occupy

$$\begin{aligned} rdH_\varphi &= rJdr \\ 2\pi rJdr &= i \\ rjdr &= \frac{i}{2\pi} \end{aligned} \tag{11}$$

Thus

$$dH_\varphi = \frac{i}{2\pi r} = \frac{i}{2\pi a_1} = (H_{\varphi 2} - H_{\varphi 1})_{a_1} \tag{12}$$

$$B\left[I_1(\gamma a_1) + \frac{I_0(\gamma a)}{K_0(\gamma a)} K_1(\gamma a_1) - I_1(\gamma a_1)\left(1 - \frac{I_0(\gamma a)K_0(\gamma a_1)}{K_0(\gamma a)I_0(\gamma a_1)}\right)\right] = \frac{i}{2\pi a_1}$$

$$B = \frac{i}{2\pi a_1}\left[\frac{I_0(\gamma a)}{K_0(\gamma a)} K_1(\gamma a_1) + \frac{I_0(\gamma a)}{K_0(\gamma a)} \frac{K_0(\gamma a_1)}{I_0(\gamma a_1)} I_1(\gamma a_1)\right]^{-1}$$

$$B = \frac{i}{2\pi a_1} \frac{K_0(\gamma a)}{I_0(\gamma a)I_1(\gamma a_1)}\left[\frac{K_1(\gamma a_1)}{I_1(\gamma a_1)} + \frac{K_0(\gamma a_1)}{I_0(\gamma a_1)}\right]^{-1} \tag{13}$$

at $r = a_1$

$$E_{z1} = E_{z2} = \left(\frac{-j\gamma}{\omega\epsilon}\right)\left(\frac{i}{2\pi a_1}\right)\frac{K_0(\gamma a)}{I_0(\gamma a)} \frac{I_0(\gamma a_1)}{I_1(\gamma a_1)}$$
$$\left(1 - \frac{I_0(\gamma a)}{K_0(\gamma a)} \frac{K_0(\gamma a_1)}{I_0(\gamma a_1)}\right)\left[\frac{K_1(\gamma a_1)}{I_1(\gamma a_1)} + \frac{K_0(\gamma a_1)}{I_0(\gamma a_1)}\right]^{-1} \tag{14}$$

Now

$$\frac{1}{\omega\epsilon} = \frac{\sqrt{\mu/\epsilon}}{\beta_0} = \frac{377}{\beta_0} \tag{15}$$

Hence

$$j\beta V = E_z = j\frac{\gamma^2}{\beta_0} I_0^2(\gamma a_1)G(\gamma a, \gamma a_1)i$$

$$V = \left(\frac{\gamma}{\beta_0}\right)\left(\frac{\gamma}{\beta}\right) I_0^2(\gamma a_1)G(\gamma a, \gamma a_1)q \tag{16}$$

$$G(\gamma a, \gamma a_1) = 60\left[\frac{K_0(\gamma a_1)}{I_0(\gamma a_1)} - \frac{K_0(\gamma a)}{I_0(\lambda a)}\right] \tag{17}$$

In obtaining this form, use was made of the fact that

$$K_1(z)I_0(z) + K_0(z)I_1(z) = \frac{1}{z}$$

Now

$$Q = \frac{\beta}{\omega C_1 (E^2/\beta^2 P)} \tag{18}$$

where $(E^2/\beta^2 P)$ is the value of this quantity at $r = a_1$. In order to evaluate Q we note that

$$V = -\frac{jV}{\omega C_1} i = \frac{-j(j\beta)}{\omega C_1} i$$

$$V = \frac{\beta}{\omega C_1} i$$

$$\frac{\beta}{\omega C_1} = \frac{V}{i} = \left(\frac{\gamma}{\beta_0}\right)\left(\frac{\gamma}{\beta}\right) I_0^2(\gamma a_1) G(\gamma a, \gamma a_1) \tag{20}$$

$$\frac{\beta}{\omega C_1} = \left(\frac{\beta}{\beta_0}\right)\left(\frac{\gamma}{\beta}\right)^2 I_0^2(\gamma a_1) G(\gamma a, \gamma a_1)$$

On the axis, $(E^2/\beta^2 P)$ has a value $(E^2/\beta^2 P)_0$

$$(E^2/\beta^2 P)_0 = \left(\frac{\beta}{\beta_0}\right)\left(\frac{\gamma}{\beta}\right)^4 F^3(a) \tag{21}$$

At a radius a_1

$$(E^2/\beta^2 P) = \left(\frac{\beta}{\beta_0}\right)\left(\frac{\gamma}{\beta}\right)^4 F^3(\gamma a) I_0^2(\gamma a_1) \tag{22}$$

Hence

$$Q(\gamma/\beta)^2 = \frac{G(\gamma a, \gamma a_1)}{F^3(\gamma a)} \tag{23}$$

APPENDIX V

DIODE EQUATIONS

FROM LLEWELLYN AND PETERSON

These apply to electrons injected into a space between two planes a and b normal to the x direction. Plan b is in the $+x$ direction from plane a. Current density I and convection current q are positive in the $-x$ direction. The d-c velocities u_a , u_b and the a-c velocities v_a , v_b are in the $+x$ direction. T is the transit time. The notation in this appendix should not be confused with that used in other parts of this book. It was felt that it would be confusing to change the notation in Llewellyn's and Peterson's[1] well-known equations.

TABLE I

ELECTRONICS EQUATIONS

Numerics Employed:

$$\eta = 10^7 \frac{e}{m} = 1.77 \times 10^{15}, \qquad \epsilon = 1/(36\pi \times 10^{11}) \frac{\eta}{\epsilon} \doteq 2 \times 10^{28}$$

Direct-Current Equations:

Potential-velocity: $\eta V_D = (1/2)u^2$ (1)

Space-charge-factor definition: $\zeta = 3(1 - T_0/T)$

Distance: $x = (1 - \zeta/3)(u_a + u_a)T/2$ (2)

Current density: $(\eta/\epsilon)I_D = (u_a + u_b)2\zeta/T^2$

Space-charge ratio: $I_D/I_m = (9/4)\zeta(1 - \zeta/3)^2$ (3)

Limiting-current density:

$$I_m = \frac{2.33}{10^6} \frac{(\sqrt{V_{Da}} + \sqrt{V_{Db}})^3}{x^2} \tag{4}$$

Alternating-Current Equations:

Symbols employed:

$$\beta = i\theta, \qquad \theta = \omega T, \qquad i = \sqrt{-1}$$

[1] F. B. Llewellyn and L. C. Peterson "Vacuum Tube Networks," *Proc. I.R.E.*, vol 32, pp. 144–166, March, 1944.

$$P = 1 - e^{-\beta} - \beta e^{-\beta} \doteq \frac{\beta^2}{2} - \frac{\beta^3}{3} + \frac{\beta^4}{8} \cdots$$

$$Q = 1 - e^{-\beta} \doteq \beta - \frac{\beta^2}{2} + \frac{\beta^3}{6} - \frac{\beta^4}{24} \cdots$$

$$S = 2 - 2e^{-\beta} - \beta - \beta e^{-\beta} \doteq -\frac{\beta^3}{6} + \frac{\beta^4}{12} - \frac{\beta^5}{40} + \frac{\beta^6}{180}$$

General equations for alternating current

q = alternating conduction-current density

v = alternating velocity

$$\left.\begin{aligned} V_b - V_a &= A^*I + B^*q_a + C^*v_a \\ q_b &= D^*I + E^*q_a + F^*v_a \\ v_b &= G^*I + H^*q_a + I^*v_a \end{aligned}\right\} \tag{5}$$

TABLE II

VALUES OF ALTERNATING-CURRENT COEFFICIENTS

$$A^* = \frac{1}{\epsilon} u_a + u_b \frac{T^2}{2} \frac{1}{\beta} \left[1 - \frac{\zeta}{3}\left(1 - \frac{12S}{\beta^3}\right)\right]$$

$$B^* = \frac{1}{\epsilon} \frac{T^2}{\beta^3} [u_a(P - \beta Q) - u_b P + \zeta(u_a + u_b)P]$$

$$C^* = -\frac{1}{\eta} 2\zeta(u_a + u_b) \frac{P}{\beta^2}$$

$$D^* = 2\zeta \frac{(u_a + u_b)}{u_b} \frac{P}{\beta^2}$$

$$E^* = \frac{1}{u_b} [u_b - \zeta(u_a + u_b)]e^{-\beta}$$

$$F^* = \frac{\epsilon}{\eta} \frac{2\zeta}{T^2} \frac{(u_a + u_b)}{u_b} \beta e^{-\beta}$$

$$G^* = -\frac{\eta}{\epsilon} \frac{T^2}{\beta^3} \frac{1}{u_b} [u_b(P - \beta Q) - u_a P + \zeta(u_a + u_b)P]$$

$$H^* = -\frac{\eta}{\epsilon} \frac{T^2}{2} \frac{(u_a + u_b)}{u_b}$$

$$(1 - \zeta) \frac{e^{-\beta}}{\beta}$$

$$I^* = \frac{1}{u_b} [u_a - \zeta(u_a + u_b)]e^{-\beta}$$

Complete space-charge, $\zeta = 1$.

$$A^* = \frac{1}{\epsilon} (u_a + u_b) \frac{T^2}{3\beta} \left(1 + \frac{6S}{\beta^3}\right)$$

$$B^* = \frac{1}{\epsilon} \frac{T^2}{\beta^3} u_a(2P - \beta Q)$$

$$C^* = -\frac{2}{\eta}(u_a + u_b)\frac{P}{\beta^2}$$

$$D^* = 2\frac{(u_a + u_b)}{(u_b)}\frac{P}{\beta^2}$$

$$E^* = -\frac{u_a}{u_b}e^{-\beta}$$

$$F^* = \frac{\epsilon}{\eta}\frac{2}{T^2}\frac{(u_a + u_b)}{(u_b)}\beta e^{-\beta}$$

$$G^* = -\frac{\eta}{\epsilon}\frac{T^2}{\beta^3}(2P - \beta Q)$$

$$H^* = 0$$

$$I^* = -e^{-\beta}$$

APPENDIX VI

EVALUATION OF IMPEDANCE AND Q FOR THIN AND SOLID BEAMS[1]

Let us first consider a thin beam whose breadth is small enough so that the field acting on the electrons is essentially constant. The normal mode solutions obtained in Chapters VI and VII apply only to this case. The more practical situation of a thick beam will be considered later. The normal mode method consists of simultaneously solving two equations, one relating the r-f field produced on the circuit by an impressed r-f current from the electron stream and the other relating r-f current produced in the electron stream by an impressed r-f field from the circuit.

We have the circuit equation

$$E = -\left[\frac{\Gamma^2\Gamma_0 K}{\Gamma^2 - \Gamma_0^2} + \frac{2jQK\Gamma^2}{\beta_e}\right] i \tag{1}$$

and the electronic equation

$$i = \frac{j\beta_e}{(j\beta_e - \Gamma)^2}\frac{I_0}{2V_0} E. \tag{2}$$

The solution of these two equations gives Γ in terms of Γ_0, K, and Q, which must be evaluated separately for the particular circuit being considered.

The field solution is obtained by solving the field equations in various regions and appropriately matching at the boundaries. For a hollow beam of electrons of radius b traveling in the z direction inside a helix of radius a and pitch angle ψ, the matching consists of finding the admittances $\left(\frac{H_\varphi}{E_z}\right)$ inside and outside the beam and setting the difference equal to the admittance of the beam. Thus the admittance just outside the beam for an idealized helix will be[2]

$$Y_0 = \frac{H_{\varphi 0}}{E_{z0}} = j\frac{\omega\epsilon}{\gamma}\frac{I_1(\gamma b) - \delta K_1(\gamma b)}{I_0(\gamma b) + \delta K_0(\gamma b)}, \tag{3}$$

[1] This appendix is taken from R. C. Fletcher, "Helix Parameters in Traveling-Wave Tube Theory," *Proc. I.R.E.*, Vol. 38, pp. 413–417 (1950).

[2] L. J. Chu and J. D. Jackson, "Field Theory of Traveling-Wave Tubes," *I.R.E., Proc.*, Vol. 36, pp. 853–863, July, 1948.

O. E. H. Rydbeck, "Theory of the Traveling-Wave Tube," *Ericsson Technics*, No. 46 pp. 3–18, 1948.

where

$$\delta = \frac{1}{K_0^2(\gamma a)}\left(\left(\frac{\beta_0 a \cot \Psi}{\gamma a}\right)^2 I_1(\gamma a)K_1(\gamma a) - I_0(\gamma a)K_0(\gamma a)\right),$$

$$\beta_0^2 = \omega^2\mu\epsilon,$$

and

$$\gamma^2 = -\Gamma^2 - \beta_0^2.$$

(The I's and K's are modified Bessel functions). The admittance inside the beam is

$$Y_i = \frac{H_{\varphi i}}{E_{zi}} = \frac{j\omega\epsilon}{\gamma}\frac{I_1(\gamma b)}{I_0(\gamma b)}. \tag{4}$$

Boundary conditions require that $E_{z0} = E_{zi} = E_z$ and $H_{z0} - H_{zi} = \frac{i}{2\pi b}$. Combining the boundary conditions, we see that

$$Y_0 - Y_i = \frac{1}{2\pi b}\frac{i}{E_z}, \tag{5}$$

where the ratio of $\frac{i}{E_z}$ is given by (2). Thus the field method gives two equations which are equivalent to the circuit and electronic equations of the normal mode method.

A6.1 Normal Mode Parameters for Thin Beam

The constants appearing in eq. (1) can be evaluated by equating the circuit equation (1) to the circuit equation (5). Thus if $Y_c = Y_0 - Y_i$,

$$-\frac{\Gamma^2 Y_0 K}{\Gamma^2 - \Gamma_0^2} - \frac{2jQK\Gamma^2}{\beta_e} = +\frac{1}{2\pi b Y_c}. \tag{6}$$

The constants can be obtained by expanding each side of eq. (6) in terms of the zero and pole occurring in the vicinity of Γ_0. Thus if γ_0 and γ_p are the zero and pole of Y_c, respectively,

$$Y_c \simeq -(\gamma_p - \gamma_0)\left(\frac{\partial Y_c}{\partial \gamma}\right)_{\gamma=\gamma_0}\left(\frac{\gamma - \gamma_0}{\gamma - \gamma_p}\right), \tag{7}$$

and the two sides of eq. (6) will be equivalent if

$$\Gamma_0^2 = -\gamma_0^2 - \beta_0^2, \tag{8}$$

$$\frac{2Q}{\beta_e} = \left(1 + \frac{\beta_0^2}{\gamma_0^2}\right)^{-1/2}\frac{\gamma_0}{\gamma_p^2 - \gamma_0^2}, \tag{9}$$

and

$$\frac{1}{K} = -j\pi b\gamma_0^2\left(1 + \frac{\beta_0^2}{\gamma_0^2}\right)^{3/2}\left(\frac{\partial Y_c}{\partial Y}\right)_{\gamma=\gamma_0}. \tag{10}$$

γ_0 and γ_p can be obtained from eqs. (3) and (4) through the implicit equations

$$(\beta a \cot \Psi)^2 = (\gamma_0 a)^2 \frac{I_0(\gamma_0 a)K_0(\gamma_0 a)}{I_1(\gamma_0 a)K_1(\gamma_0 a)}, \tag{11}$$

$$\frac{I_0(\gamma_p b)}{K_0(\gamma_p b)} = -\frac{1}{K_0^2(\gamma_p a)} \cdot \left[\left(\frac{\beta_0 a \cot \Psi}{\gamma_p a}\right)^2 I_1(\gamma_p a)K_1(\gamma_p a) - I_0(\gamma_p a)K_0(\gamma_p a)\right], \tag{12}$$

and $1/K$ is found to be

$$\frac{1}{K} = \pi\sqrt{\frac{\epsilon}{\mu}}\left(1 + \frac{\beta_0^2}{\gamma_0^2}\right)^{3/2}\frac{\beta_0^2}{I_0^2(\gamma_0 b)}\frac{I_0(\gamma_0 a)}{K_0(\gamma_0 a)}\left[\frac{I_1(\gamma_0 a)}{I_0(\gamma_0 a)} - \frac{I_0(\gamma_0 a)}{I_1(\gamma_0 a)} + \frac{K_0(\gamma_0 a)}{K_1(\gamma_0 a)} - \frac{K_1(\gamma_0 a)}{K_0(\gamma_0 a)} + \frac{4}{\gamma_0 a}\right]. \tag{13}$$

The equations for γ_0 and K are the same as those given by Appendix II, evaluated by solving the field equations for the helix without electrons present. The evaluation of γ_p, and thus Q, represents a new contribution. Values of $Q\frac{\gamma_0}{\beta_e}\left(1 + \frac{\beta_0^2}{\gamma_0^2}\right)^{-1/2}$ are plotted in Fig. A6.1 as a function of $\gamma_0 a$ for various ratios of b/a. (It should be noted that for most practical applications the factor $\frac{\gamma_0}{\beta_e}\left(1 + \frac{\beta_0^2}{\gamma_0^2}\right)^{-1/2}$ is very close to unity, so that the ordinate is practically the value of Q itself.)

Appendix IV gives a method for estimating Q based on the solution of the field equations for a conductor replacing the helix and considering the resultant field to be $-\frac{2jKQI^2}{\beta_e} i$. This estimate of Q is plotted as the dashed lines of Fig. A6.1.

A6.2 Thick Beam Case

For an electron beam which entirely fills the space out to the radius b, the electronic equations of both the normal mode method and the field method are altered in such a way as to considerably complicate the solution. In order to find a solution for this case some simplifying assumptions must be made. A convenient type of assumption is to replace the thick beam by an "equivalent" thin beam, for which the solutions have already been worked out.

Two beams will be equivalent if the value of $\frac{H_\varphi}{E_z}$ is the same outside the beams, since the matching to the circuit depends only on this admittance.

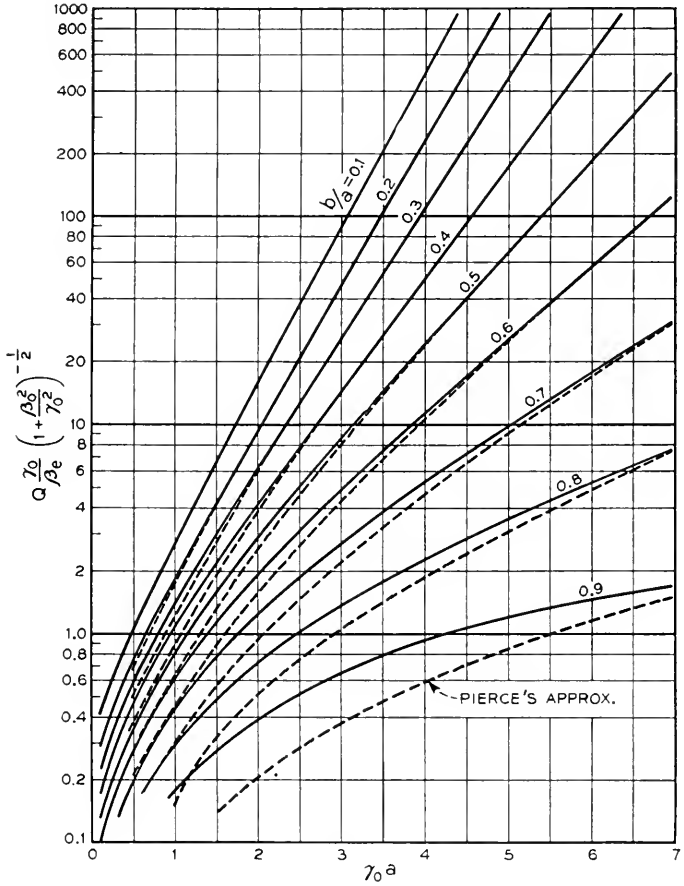

Fig. A6.1—Passive mode parameter Q for a hollow beam of electrons of radius b inside a helix of radius a and natural propagation constant γ_0. The solid line was obtained by equating the circuit equation of the normal mode method, which defines Q, with a corresponding circuit equation found from the field theory method. The dashed line was obtained in Appendix IV from a solution of the field equations for a conductor replacing the helix.

The problem, then, of making a thin beam the equivalent of a thick beam is the problem of arranging the position and current of a thin beam to give the same admittance at the radius b of the thick beam. This is of course impossible for all values of γ. It is desirable therefore that the admittances

be the same close to the complex values of γ which will eventually solve the equations.

The solution of the field equations for the solid beam yields the value for $\frac{H_\varphi^{(1)}}{E_z}$ at the radius b as

$$\frac{H_\varphi}{E_z} = \frac{j\omega\epsilon}{\gamma} \frac{nI_1(n\gamma b)}{I_0(n\gamma b)}, \tag{14}$$

where

$$n^2 = 1 + \frac{1}{\beta_0} \sqrt{\frac{\mu}{\epsilon}} \frac{\beta_e I_0}{2\pi b^2 V_0} \frac{1}{(j\beta_e - \Gamma)^2}. \tag{15}$$

Thus the electronic equation for the solid beam which must be solved simultaneously with the circuit equation (given above by either the normal mode approximation or the field solution) must be

$$Y_e = \frac{H_\varphi}{E_z} - Y_i = \frac{j\omega\epsilon b}{\gamma b} \left[\frac{nI_1(n\gamma b)}{I_0(n\gamma b)} - \frac{I_1(\gamma b)}{I_0(\gamma b)} \right]. \tag{16}$$

Complex roots for γ will be expected in the vicinity of real values of γ for which $Y_e \approx Y_c$ and $\frac{dY_e}{d\gamma} \approx \frac{dY_c}{d\gamma}$. By plotting Y_e and Y_c vs. real values of γ, it is found that the two curves become tangent close to the value of γ for which $n = 0$, using typical operating conditions (Fig. A6.2). Our procedure for choosing a hollow beam equivalent of the solid beam, then, will be to equate the values of Y_e and $\frac{dY_e}{d\gamma}$ at $n = 0$. This will give us two equations from which to solve for the electron beam diameter and d-c current for the equivalent hollow beam.

If the hollow beam is placed at the radius sb with a current of tI_0, the value of $\frac{H_\varphi}{E_z}$ at the radius b gives the value for Y_{eH} as

$$\begin{aligned} Y_{eH} = \left(\frac{H_\varphi}{E_z}\right)_b - Y_i = -j\omega\epsilon b \frac{t}{2}(1 - n^2) \frac{I_0^2(s\gamma b)}{I_0^2(\gamma b)} \\ \cdot \left(1 - \gamma^2 b^2 I_0^2(s\gamma b) \frac{t}{2}(1 - n^2) \left[\frac{K_0(s\gamma b)}{I_0(s\gamma b)} - \frac{K_0(\gamma b)}{I_0(\gamma b)} \right] \right)^{-1}. \end{aligned} \tag{17}$$

Equating this with eq. (16) at $n = 0$ yields the equation

$$\frac{1}{t} = \frac{1}{2} \theta^2 I_0^2(s\theta) \left[\frac{K_0(s\theta)}{I_0(s\theta)} + \frac{K_1(\theta)}{I_1(\theta)} \right], \tag{18}$$

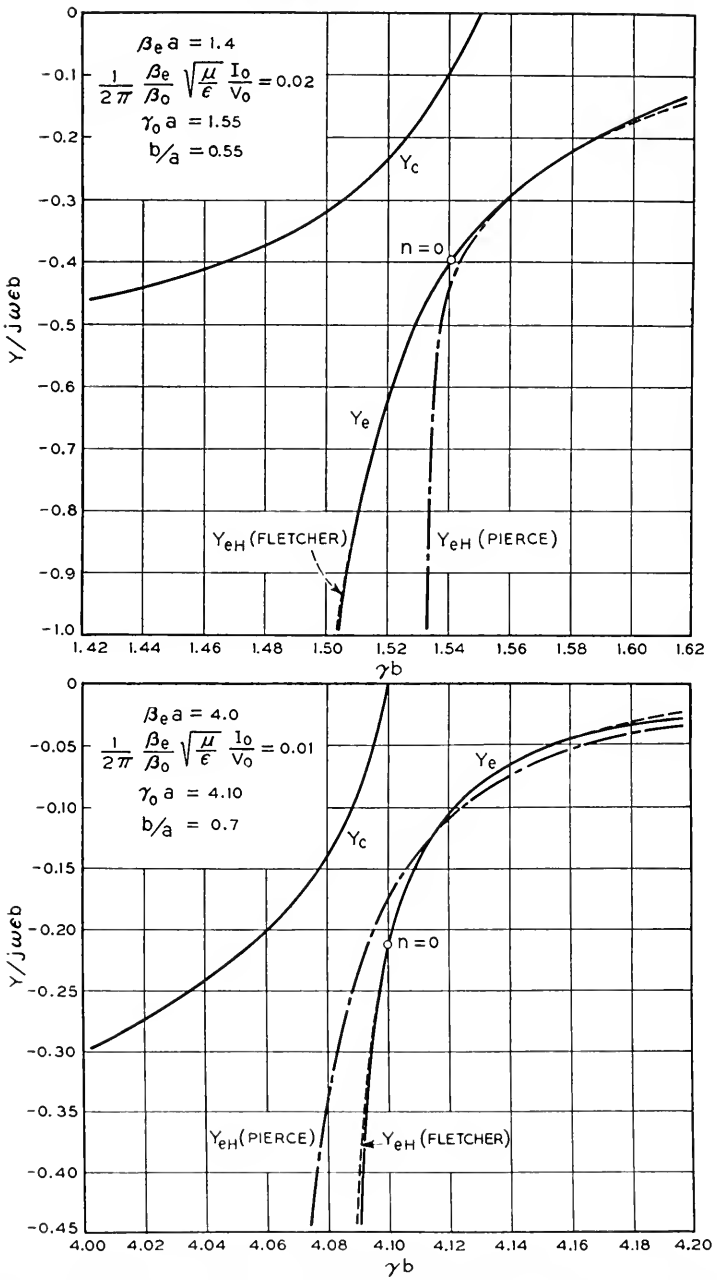

Fig. A6.2—Electronic admittance Y_e of a solid electron beam of radius b and circuit admittance Y_c of a helix of radius a plotted vs. real values of the propagation constant γ in the vicinity of where $\frac{dY_e}{d\gamma} = \frac{dY_c}{d\gamma}$ where complex solutions for γ are expected, for two typical sets of operating conditions. Plotted on the same graph is the electron admittance Y_{eH} for two equivalent hollow electron beams: the dashed curve (Fletcher) is matched to Y_e at $n = 0$, while the dot-dashed curve (Pierce, Appendix IV) is matched at $n = 1$ (off the graph).

where $\theta = \gamma_e b$ and γ_e is the value of γ at $n = 0$; i.e. for $\gamma_e \gg \beta_0$

$$\gamma_e = \beta_e + \sqrt{\frac{1}{\beta_0}\sqrt{\frac{\mu}{\epsilon}}\frac{\beta_e I_0}{2\pi b^2 V_0}} \approx \beta_e . \tag{19}$$

In the vicinity of $n = 0$, n varies very rapidly with γ, and hence matching $\left(\frac{\partial Y_e}{\partial n}\right)_\gamma$ is practically the same as matching $\frac{dY_e}{d\gamma}$. With this approximation eqs. (16) and (17) can be differentiated with respect to n and set equal at

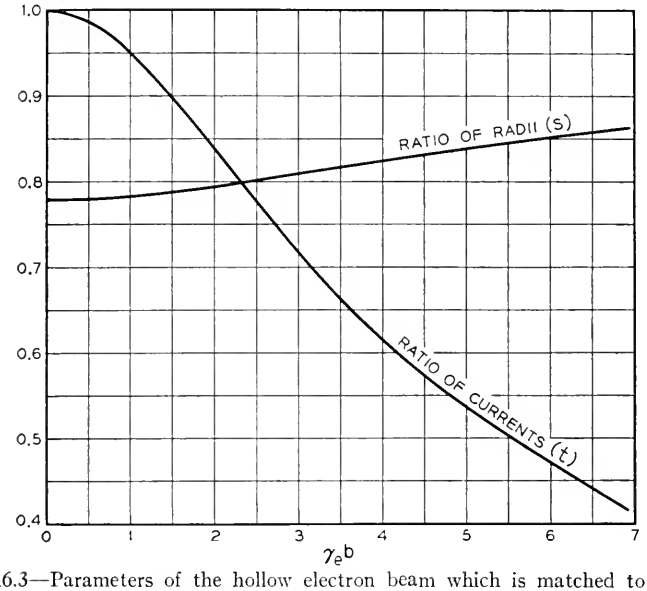

Fig. A6.3—Parameters of the hollow electron beam which is matched to the solid electron beam of radius b and current I_0 at $\gamma = \gamma_e \simeq \beta_e$, where $n = 0$. sb is the radius and tI_0 is the current of the equivalent hollow beam.

$n = 0$ to yield the second relation

$$\frac{1}{t} = \theta^2 I_0^2(\theta) I_0^2(s\theta)\left[\frac{K_0(s\theta)}{I_0(s\theta)} + \frac{K_1(\theta)}{I_1(\theta)}\right]^2 \tag{20}$$

Equations (18) and (20) can then be solved to give the implicit equation for s as

$$\frac{K_0(s\theta)}{I_0(s\theta)} = -\frac{K_1(\theta)}{I_1(\theta)} + \frac{1}{2I_1^2(\theta)} \tag{21}$$

and the simpler equation for t

$$t = \frac{4}{\theta^2}\frac{I_1^2(\theta)}{I_0^2(s\theta)} . \tag{22}$$

s and t are plotted as a function of θ in Fig. A6.3. The value of Y_{eH} using these values of s and t is compared in Fig. A6.2 with Y_e in the vicinity of where Y_c is almost tangent to Y_e for two typical sets of operating conditions.

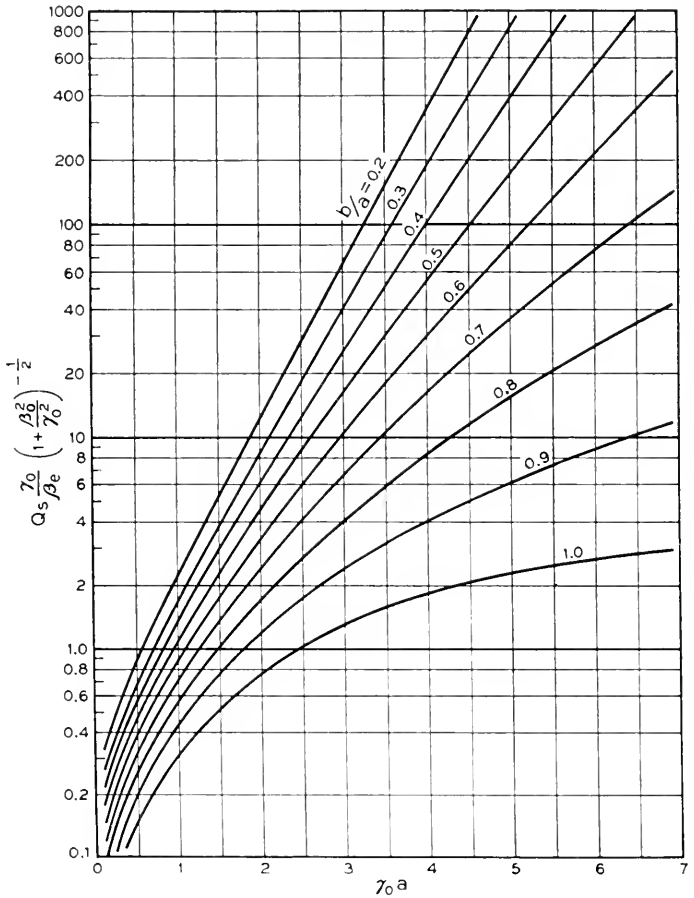

Fig. A6.4—Passive mode parameter Q_s for a solid beam of electrons of radius b inside a helix of radius a and natural propagation constant γ_0, obtained from the equivalent hollow beam parameters of Fig. 3 taken at $\gamma_e = \gamma_0$. All the normal mode solutions which have been found[2], [3] for a hollow beam will be approximately valid for a solid beam if Q is replaced by Q_s and K is replaced by K_s (Fig. 5).

It is of course possible to pick other criteria for determining an "equivalent" hollow beam. In Chapter XIV, in essence, Y_e and Y_{eH} were expanded in terms of $(1 - n^2)$ and the coefficients of the first two terms were equated. This has been done for the cylindrical beams, and the values of s and t found by this method determine values of Y_{eH} shown in Fig. A6.2. The greater

departure from the true curve of V_e would indicate that this approximation is not as good as that described above.

It is now possible to find the values of Q_s and K_s appropriate to the solid

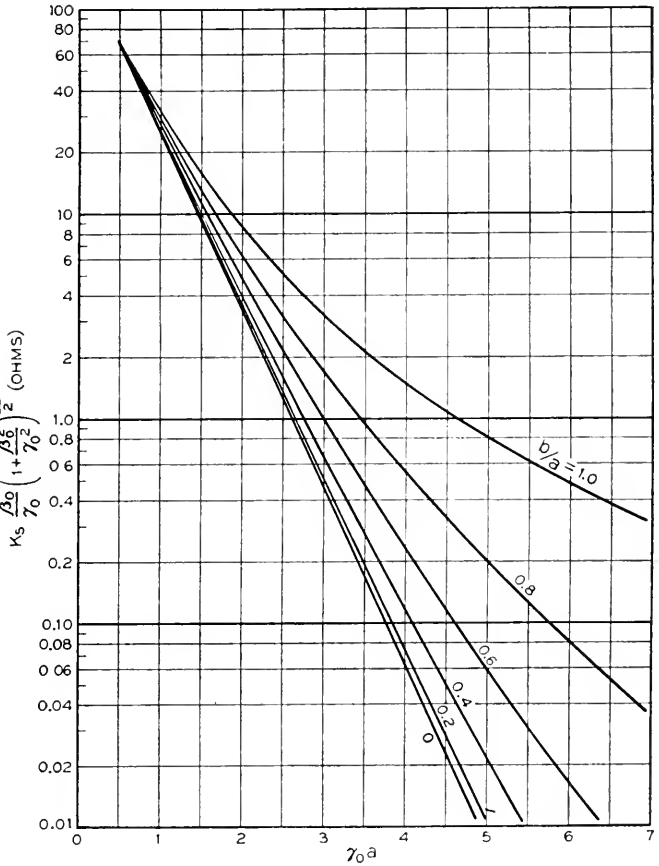

Fig. A6.5—Circuit impedance K_s for a solid beam of electrons of radius b inside a helix of radius a and natural propagation constant γ_0, obtained from the equivalent hollow beam parameters of Fig. 3 taken at $\gamma_e = \gamma_0$. K_s should replace $K = \frac{E^{2\,(2),\,(3)}}{2\beta^2_p}$ in order for the normal mode solutions for a hollow beam to be applicable to a solid beam.

beam. Thus if $Q\left(\gamma_0 a, \frac{b}{a}\right)$ and $K\left(\gamma_0 a, \frac{b}{a}\right)$ are the values for the hollow beam calculated from eqs. (9), (12) and (13),

$$Q_s = Q\left(\gamma_0 a, s\,\frac{b}{a}\right), \tag{23}$$

and

$$K_s = tK\left(\gamma_0 a,\ s\frac{b}{a}\right). \tag{24}$$

The t is placed in front of K in eq. (24) because tI_0 and K appear in the thin beam solutions only in the combination tI_0K. Using tK instead of K allows us to use I_0, the actual value of the current in the solid beam in the solutions instead of tI_0, the equivalent current. Values of $Q_s \frac{\gamma_0}{\beta_e}\left(1+\frac{\beta_0^2}{\gamma_0^2}\right)^{-1/2}$ and $K_s \frac{\beta_0}{\gamma_0}\cdot\left(1+\frac{\beta_0^2}{\gamma_0^2}\right)^{+3/2}$ are plotted vs. $\gamma_0 a$ in Figs. A6.4 and A6.5 for different values of b/a and for values of t and s taken at $\gamma_e = \gamma_0$. All the solutions obtained for the hollow beam will be valid for the solid beam if Q_s and K_s are substituted for Q and K.

APPENDIX VII

HOW TO CALCULATE THE GAIN OF A TRAVELING-WAVE TUBE

The gain calculation presented here neglects the effect at the output of all waves except the increasing wave. Thus, it can be expected to be accurate only for tubes with a considerable net gain. The gain is expressed in db as

$$G = A + BCN \tag{1}$$

Here A represents an initial loss in setting up the increasing wave and BCN represents the gain of the increasing wave.

We will modify (1) to take into account approximately the effect of the cold loss of L db in reducing the gain of the increasing wave by writing

$$G = A + [BCN - \alpha L] \tag{2}$$

Here α is the fraction of the cold loss which should be subtracted from the gain of the increasing wave. This expression should hold even for moderately non-uniform loss (see Fig. 9.5).

Thus, what we need to know to calculate the gain are the quantities

$$A, B, C, N, \alpha, L$$

A7.1 Cold Loss L db

The best way to get the cold loss L is to measure it. One must be sure that the loss measured is the loss of a wave traveling in the circuit and not loss at the input and output couplings.

A7.2 Length of Circuit in Wavelengths, N

We can arrive at this in several ways. The ratio of the speed of light c to the speed of an electron u_0 is

$$\frac{c}{u_0} = \frac{505}{\sqrt{V_0}} \tag{3}$$

where V_0 is the accelerating voltage. Thus, if ℓ is the length of the circuit and λ is the free-space wavelength and λ_g is the wavelength along the axis of

the helix

$$\lambda_g = \lambda \frac{u_0}{c} \tag{4}$$

$$N = \frac{\ell}{\lambda_g} = \frac{\ell}{\gamma} \frac{c}{u_0} \tag{5}$$

Also, if $\mathcal{L}_w$ is the total length of wire in the helix, approximately

$$N = \frac{L_w}{\lambda} \tag{6}$$

A7.3 The Gain Parameter C

The gain parameter can be expressed

$$C = \left(\frac{E^2}{\beta^2 P} \frac{I_0}{8V_0} \right)^{1/3} = \left(\frac{KI_0}{4V_0} \right)^{1/3} \tag{7}$$

Here K is the helix impedance properly defined. I_0 is the beam current in amperes and V_0 is the beam voltage.

A7.4 Helix Impedance K

In Fig. 5 of Appendix VI, $K \left(\frac{\beta_0}{\gamma_0} \right) \left(1 + \left(\frac{\beta_0}{\gamma_0} \right)^2 \right)^{3/2}$ is plotted vs. $\gamma_0 a$ for values of b/a. K_s is the effective value of K for a solid beam of radius b, and a is the radius of the helix. γ_0 is to be identified with γ for present purposes, and is given by

$$\gamma_0 = \frac{2\pi}{\lambda_g} \left[1 - \left(\frac{\gamma_g}{\lambda} \right)^2 \right]^{1/2} \tag{8}$$

where λ_g is given in terms of λ by (4). We see that in most cases (for voltages up to several thousand)

$$(\lambda_g/\lambda)^2 \ll 1 \tag{9}$$

and we may usually use as a valid approximation

$$\gamma_0 = \frac{2\pi}{\lambda_g} \tag{10}$$

and

$$\gamma_0 a = \frac{2\pi a}{\lambda_g} \tag{11}$$

As $\beta_0 = 2\pi/\lambda$, this approximation gives

$$1 + \left(\frac{\beta_0}{\gamma_0} \right)^2 = 1 + \left(\frac{\lambda_g}{\lambda} \right)^2$$

and we may assume

$$\left(1 + \left(\frac{\beta_0}{\gamma_0}\right)^2\right)^{3/2} = 1 \tag{12}$$

Thus, we may take K_s as the ordinate of Fig. 5 multiplied by c/u_0, from (3), for instance.

The true impedance may be somewhat less than the impedance for a helically conducting sheet. If the ratio of the circuit impedance to that of a helically conducting sheet is known (see Sections 3 and 4.1 of Chapter III, and Fig. 3.13, for instance), the value of K_s from Fig. 5 can be multiplied by this ratio.

A7.5 The Space-Charge Parameter Q

The ordinate of Fig. 4 of Appendix VI shows $Q_s \frac{\gamma_0}{\beta_e}\left(1 + \left(\frac{\beta_0}{\gamma_0}\right)^2\right)^{-1/2}$ vs. γa for several values of b/a. Here Q_s is the effective value of Q for a solid beam of radius b. As before, for beam voltages of a few thousand or lower, we may take

$$\left(1 + \left(\frac{\beta_0}{\gamma_0}\right)^2\right)^{-1/2} = 1$$

The quantity β_e is just

$$\beta_e = \frac{2\pi a}{\lambda_g} \tag{13}$$

and from (8) we see that for low beam voltages we can take

$$\beta_e = \gamma = \gamma_0$$

so that the ordinate in Fig. 4 can usually be taken as simply Q_s.

A7.6 The Increasing Wave Parameter B

In Fig. 8.10, B is plotted vs. QC. C can be obtained by means of Sections 3 and 4, and Q by means of Section 5. Hence we can obtain B.

A7.7 The Gain Reduction Parameter α

From (2) we see that we should subtract from the gain of the increasing wave in db α times the cold loss L in db. In Fig. 8.13 a quantity $\partial x_1/\partial d$, which we can identity as α, is plotted vs. QC.

A7.8 The Loss Parameter d

The loss parameter d can be expressed in terms of the cold loss, L in db,

the length of the circuit in wavelengths, N, and C

$$d = \left(\frac{2.3L}{20}\right)\left(\frac{1}{2\pi\sqrt{NC}}\right) \tag{14}$$

$$d = 0.0183\,\frac{L}{\sqrt{NC}} \tag{15}$$

A7.9 The Initial Loss A

The quantity A of (2) is plotted vs. d in Fig. 9.3. This plot assumes $QC = 0$, and may be somewhat in error. Perhaps Fig. 9.4 can be used in estimating a correction; it looks as if the initial loss should be less with $QC \neq 0$ even when $d \neq 0$. In any event, an error in A means only a few db, and is likely to make less error in the computed gain than does an error in B, for instance.

Technical Publications by Bell System Authors Other Than in the Bell System Technical Journal

*Progress in Coaxial Telephone and Television Systems.** L. G. ABRAHAM.[1] *A.I.E.E., Trans.*, V. 67, pt. 2, pp. 1520–1527, 1948.

ABSTRACT—This paper describes coaxial systems used in the Bell System to transmit telephone and television signals. Development of this system was started some time ago, with systems working before the war between New York and Philadelphia and later between Minneapolis, Minnesota and Stevens Point, Wisconsin. Various stages in the progress of this development have been described in previous papers and the telephone terminal equipment has been recently described. This paper will outline how the system works and discuss some transmission problems, leaving a complete technical description for a number of later papers.

Use of the Relay Digital Computer. E. G. ANDREWS and H. W. BODE.[1] *Elec. Engg.*, V. 69, pp. 158–163, Feb., 1950.

ABSTRACT—This paper is concerned primarily with the operating features of the computer and its application to problems of scientific and engineering interest. The material herein has been derived largely from the experience gained with one of the computers during a trial period of about 5 months before final delivery. An effort was made during that time to try the machine out on a variety of difficult computing problems of varying character to obtain experience in its operation and to establish as well as possible what its range of usefulness might be.

Longitudinal Noise in Audio Circuits. H. W. AUGUSTADT and W. F. KANNENBERG.[1] *Audio Engg.*, V. 34, pp. 18–19, Feb., 1950.

ABSTRACT—The words "longitudinal interference" have often been used to explain the origin of unknown noise in audio circuits with little actual regard to the source of the interference. In this respect, the usage of these words is similar to the popular usage of the word "gremlins". We attribute to gremlins troubles whose causes are unknown without much attempt to delve deeper into the matter. Similarly in the audio facilities field, many noise troubles are attributed to "longitudinal interference" or "longitudinals" or even simply "line noise" without a clear understanding of the nature of the trouble or the actual meaning of the terms. The noise trouble, however, still persists irrespective of the name applied to it until its causes are thoroughly understood and the correct remedial action is applied. This

*A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

paper describes and illustrates, with representative examples, various types of common noise induction in order to lead to an understanding of their nature. The paper includes, in addition, a discussion of simple remedies which may be employed for representative cases of noise troubles due to longitudinal induction.

Mobile Radio. A. BAILEY.[3] *A.I.E.E., Trans.*, V. 67, pt. 2, pp. 923–931, 1948.

*Stabilized Permanent Magnets.** P. P. CIOFFI.[1] *A.I.E.E., Trans.*, V. 67, pt. 2, pp. 1540–1543, 1948.

ABSTRACT—Permanent magnets are stabilized against forces tending to demagnetize them, by partial demagnetization. It is shown that, after such stabilization, the magnet operates at a point on a secondary demagnetization curve. This curve may be treated identically as the major demagnetization curve is treated in ordinary magnet design problems. Formulas are developed for determining secondary demagnetization curves from the major demagnetization curve when stabilization is achieved by magnetization of the magnet before assembly, and by an applied magnetomotive force after magnetization in assembly.

It will be shown that, when the magnet is partially demagnetized for the purpose of stabilization, its operating point lies on a curve which, for convenience, will be called a secondary demagnetization curve. The object of this paper is to discuss the derivation of secondary demagnetization curves for given conditions of stability against demagnetizing forces and their applications to magnet design problems.

*Relay Preference Lockout Circuits in Telephone Switching.** A. E. JOEL, JR.[1] *A.I.E.E., Trans.*, V. 67, pt. 2, pp. 1720–1725, 1948.

ABSTRACT—Occasions arise in telephone switching, particularly at common controlled stages, where calls compete for the use of equipment components or switching linkages. These call requests for service are received at random by circuits which must choose among and serve them on a one-at-a-time basis. Circuits which perform this function are known as "preference lockouts". Extensive use has been made of these circuits in manual, panel, and crossbar switching systems. This paper describes the design philosophies of relay preference lockout circuits based on some of these applications.

Piezoelectric Crystals and Their Application to Ultrasonics. W. P. MASON.[1] *Book*, New York, Van Nostrand, 508 pages, 1950.

*Television Terminals for Coaxial Systems.** L. W. MORRISON, JR.[1] *Elec. Engg.*, V. 69, pp. 109–115, February, 1950.

* A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[4] A. T. & T.

ABSTRACT—The broad features of operation of the L1 Coaxial System for the transmission of television have been discussed in a recent paper (L. G. Abraham, "Progress in Coaxial Telephone and Television Systems", AIEE Transactions, Vol. 67, pp. 1520–1527, 1948). It is the purpose of this paper to describe, in somewhat more detail, the factors influencing the design of the coaxial television terminals and the features of the equipment now in service in the Bell System's Television Network. The television terminals here described were placed into network service in 1947, but in basic form are similar to experimental models developed prior to the war and used in early television transmission studies over the coaxial cable.

Alternate to Lead Sheath for Telephone Cables. A. PAONE.[3] *Corrosion*, V. 6, pp. 46–50, February, 1950.

*Bridge Erosion in Electrical Contacts and Its Prevention.** W. G. PFANN.[1] *A.I.E.E., Trans.*, V. 67, pt. 2, pp. 1528–1533, 1948.

ABSTRACT—The size of the molten bridge which forms as two contacts separate depends upon the contact material and the current. The molten bridge has two diameters, one in each contact. By pairing dissimilar contact materials an asymmetric bridge is created, in which the bridge diameters are unequal and with which is associated a self-limiting transfer tendency. Under certain conditions the use of unlike pairs can prevent the continued transfer of material from one contact to the other.

*Chess-playing Machine.** C. E. SHANNON.[1] *Sci. Am.*, V. 182, pp. 48–51, February, 1950.

*Military Teletypewriter Systems of World War II.** F. J. SINGER.[1] Bibliography. *A.I.E.E., Trans.*, V. 67, pt. 2, pp. 1398–1408, 1948.

ABSTRACT—This paper reviews the evolution of military teletypewriter communications since 1941 and briefly describes some of the important systems that were developed during the war by Bell Telephone System engineers for the armed forces.

*Optimum Coaxial Diameters.** P. H. SMITH.[1] *Electronics*, V. 23, pp. 111–112, 114, February, 1950.

ABSTRACT—The derivation of the optimum ratios is briefly described and optimum values are indicated to one part in ten thousand. In all cases the medium between conductors is assumed to be a gas with a dielectric constant approaching unity, and any effect of inner conductor supports upon the optimum conductor diameter ratio for a given property has been neglected.

*General Review of Linear Varying Parameter and Nonlinear Circuit Analysis.** W. R. BENNETT.[1] *I.R.E., Proc.*, V. 38, pp. 259–263, March, 1950.

*A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[3] A. T. & T.

ABSTRACT—Variable and nonlinear systems are classified from the standpoint of their significance in communication problems. Methods of solution are reviewed and appropriate references are cited. The paper is a synopsis of a talk given at the Symposium on Network Theory of the 1949 National I.R.E. Convention.

Some Early Long Distance Lines in the Far West. W. BLACKFORD, SR.[4] and J. F. HUTTON.[4] *Bell Tel. Mag.*, V. 28, pp. 227–237, Winter, 1949–50.

*Radio Propagation Variations at VHF and UHF.** K. BULLINGTON.[1] *I.R.E., Proc.*, V. 38, pp. 27–32, January, 1950.

ABSTRACT—The variations of received signal with location (shadow losses) and with time (fading) greatly affect both the usable service area and the required geographical separation between co-channel stations. An empirical method is given for estimating the magnitude of these variations at vhf and uhf. These data indicate that the required separation between co-channel stations is from 3 to 10 times the average radius of the usable coverage area, and depends on the type of service and on the degree of reliability required. The application of this method is illustrated by examples in the mobile radiotelephone field.

*Speaking Machine of Wolfgang von Kempelen.** H. DUDLEY[1] and T. H. TARNOCZY. *Acoustical Soc. Am., Jl.*, V. 22, pp. 151–166, March, 1950.

Perception of Speech and Its Relation to Telephony. H. FLETCHER[1] and R. H. GALT.[1] *Acoustical Soc. Am., Jl.*, V. 22, pp. 89–151, March, 1950.

ABSTRACT—This paper deals with the interpretation aspect and how it is affected when speech is transmitted through various kinds of telephone systems.

Vacuum Fusion Furnace for Analysis of Gases in Metals. W. G. GULDNER[1] and A. L. BEACH.[1] *Anal. Chem.*, V. 22, pp. 366–367, February, 1950.

Complex Stressing of Polyethylene. I. L. HOPKINS,[1] W. O. BAKER[1] and J. B. HOWARD.[1] *Jl. Applied Phys.*, V. 21, pp. 206–213, March, 1950.

Noise Considerations in Sound-Recording Transmission Systems. F. L. HOPPER.[2] References. *S.M.P.E., Jl.*, V. 54, pp. 129–139, February, 1950.

*Radiation Characteristics of Conical Horn Antennas.** A. P. KING.[1] *I.R.E., Proc.*, V. 38, pp. 249–251, March, 1950.

ABSTRACT—This paper reports the measured radiation characteristics of conical horns employing waveguide excitation. The experimentally derived gains are in excellent agreement with the theoretical results (unpublished) obtained by Gray and Schelkunoff.

The gain and effective area is given for conical horns of arbitrary proportions and the radiation patterns are included for horns of optimum design.

* A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[2] W. E. Co.
[4] Pac. T. & T.

All dimensional data have been normalized in terms of wavelength, and are presented in convenient nomographic form.

Microwaves and Sound. W. E. KOCK.[1] *Physics Today*, V. 3, pp. 20–25, March, 1950.

ABSTRACT—A recent development shows that obstacle arrays, modeled after the periodic structure of crystals, refract and focus not only electromagnetic waves, but sound waves as well. The behavior of periodic structures can be investigated by microwave and acoustic experiments on such models.

Interference Characteristics of Pulse-Time Modulation. E. R. KRETZMER.[1] *I.R.E., Proc.*, V. 38, pp. 252–255, March, 1950.

ABSTRACT—The interference characteristics of pulse-time modulation are analyzed mathematically and experimentally; particular forms examined are pulse-duration and pulse-position modulation. Both two-station and two-path interference are considered. Two-station interference is found to be characterized by virtually complete predominance of the stronger signal, and by noise of random character. Two-path interference, in the case of single-channel pulse-duration modulation, generally permits fairly good reception of speech and music signals.

*Electron Bombardment Conductivity in Diamond.** K. G. MCKAY.[1] *Phys. Rev.*, V. 77, pp. 816–825, March 15, 1950.

*Perception of Television Random Noise.** P. MERTZ.[1] References. *S.M.P.E., Jl.*, V. 54, pp. 8–34, January, 1950.

ABSTRACT—The perception of random noise in television has been clarified by studying its analogy to graininess in photography. In a television image the individual random noise grains are assumed analogous to photographic grains. Effective random noise power is obtained by cumulating and weighting actual noise powers over the video frequencies with a weighting function diminishing from unity toward increasing frequencies. These check reasonably well with preliminary experiments. The paper includes an analysis of the effect of changing the tone rendering and contrast of the television image.

*Loudness Patterns—A New Approach.** W. A. MUNSON[1] and M. B. GARDNER.[1] *Acoustical Soc. Am., Jl.*, V. 22, pp. 177–190, March, 1950.

Bell System Participation in the Work of the A.S.A. H. S. OSBORNE.[3] *Bell Tel. Mag.*, V. 28, pp. 181–190, Winter, 1949–50.

*New Electronic Telegraph Regenerative Repeater.** B. OSTENDORF, JR.[1] *Elec. Engg.*, V. 69, pp. 237–240, March, 1950.

Correlation of Gieger Counter and Hall Effect Measurements in Alloys Con-

*A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.
[3] A. T. & T.

*taining Germanium and Radioactive Antimony 124.** G. L. PEARSON,[1] J. D. STRUTHERS,[1] and H. C. THEURER.[1] *Phys. Rev.*, V. 77, pp. 809–813, March 15, 1950.

*Optical Method for Measuring the Stress in Glass Bulbs.** W. T. READ.[1] *Applied Phys., Jl.*, V. 21, pp. 250–257, March, 1950.

Programming a Computer for Playing Chess. C. E. SHANNON.[1] References. *Phil. Mag.*, V. 41, pp. 256–275, March, 1950.

ABSTRACT—This paper is concerned with the problem of constructing a program for a modern electronic computer of the EDVAC type which will enable it to play chess. Although perhaps of no practical importance the question is of theoretical interest, and it is hoped that a satisfactory solution of this problem will act as a kind of wedge in attacking other problems of a similar nature and of greater significance.

Recent Developments in Communication Theory. C. E. SHANNON.[1] *Electronics*, V. 32, pp. 80–83, April, 1950.

ABSTRACT—In this paper the highlights of this recent work will be described with as little mathematics as possible. Since the subject is essentially a mathematical one, this necessitates a sacrifice of rigor; for more precise treatments the reader may consult the references.

A Symmetrical Notation for Numbers. C. E. SHANNON.[1] *Am. Math. Monthly*, V. 57, pp. 90–93, February, 1950.

*Capacity of a Pair of Insulated Wires.** W. H. WISE.[1] *Quart. Applied Math.*, V. 7, pp. 432–436, January, 1950.

*Echoes in Transmission at 450 Megacycles from Land-to-Car Radio Units.** W. R. YOUNG, JR.[1] and L. Y. LACY.[1] *I.R.E., Proc.*, V. 38, pp. 255–258, March, 1950.

*Simplified Derivation of Linear Least Square Smoothing and Prediction Theory.** H. W. BODE[1] and C. E. SHANNON.[1] *I.R.E., Proc.*, V. 38, pp. 417–425, April, 1950.

ABSTRACT—In this paper the chief results of smoothing theory will be developed by a new method which, while not as rigorous or general as the methods of Wiener and Kolmogoroff, has the advantage of greater simplicity, particularly for readers with a background of electric circuit theory. The mathematical steps in the present derivation have, for the most part, a direct physical interpretation, which enables one to see intuitively what the mathematics is doing.

*Helix Parameters Used in Traveling Wave-Tube Theory.** R. C. FLETCHER.[1] *I.R.E., Proc.*, V. 38, pp. 413–417, April, 1950.

ABSTRACT—Helix parameters used in the normal mode solution of the traveling-wave tube are evaluated by comparison with the field equations

* A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

for a thin electron beam. Corresponding parameters for a thick electron beam are found by finding a thin beam with approximately the same r-f admittance.

*Effect of Change of Scale on Sintering Phenomena.** C. HERRING.[1] *Jl., Applied Phys.*, V. 21, pp. 301–303, April, 1950.

ABSTRACT—It is shown that when certain plausible assumptions are fulfilled simple scaling laws govern the times required to produce, by sintering at a given temperature, geometrically similar changes in two or more systems of solid particles which are identical geometrically except for a difference of scale. It is suggested that experimental studies of the effect of such a change of scale may prove valuable in identifying the predominant mechanism responsible for sintering under any particular set of conditions, and may also help to decide certain fundamental questions in fields such as creep and crystal growth.

*Mode Conversion Losses in Transmission of Circular Electric Waves Through Slightly Non-Cylindrical Guides.** S. P. MORGAN, JR.[1] *Jl., Applied Phys.*, V. 21, pp. 329–338, April, 1950.

ABSTRACT—A general expression is derived for the effective attenuation of circular electric (TE_{01}) waves owing to mode conversions in a section of wave guide whose shape deviates slightly in any specified manner from a perfect circular cylinder. Numerical results are in good agreement with experiment for the special case of transmission through an elliptically deformed section of pipe. The case of random distortions in a long wave guide line is analyzed and it is calculated, under certain simplifying assumptions, that mode conversions in a 4.732-inch copper pipe whose radius deviates by 1 mil rms from that of an average cylinder will increase the attenuation of the TE_{01} mode at 3.2 cm by an amount equal to 20% of the theoretical copper losses. The dependence on frequency of mode conversion losses in such a guide is discussed.

Acoustical Designing in Architecture. C. M. HARRIS[1] and V. O. KNUDSEN. *Book*, New York, John Wiley & Sons, Inc., 450 pages, 1950.

ABSTRACT—This book is intended as a practical guide to good acoustical designing in architecture. It is written primarily for architects, students of architecture, and all others who wish a non-mathematical but comprehensive treatise on this subject. Useful design data have been presented in such a manner that the text can serve as a convenient handbook in the solution of most problems encountered in architectural acoustics.

* A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

Contributors to this Issue

R. V. L. Hartley, A.B., Utah, 1909; B.A., Oxford, 1912; B.Sc., 1913; Instructor in Physics, Nevada, 1909–10. Engineering Department, Bell Telephone Laboratories, 1913–50. Mr. Hartley took part in the early radio telephone experiments and was thereafter associated with research on telephony and telegraphy at voice and carrier frequencies. Later, as Research Consultant he was concerned with general circuit problems. Mr. Hartley is now retired from active service.

J. R. Pierce, B.S., in Electrical Engineering, California Institute of Technology, 1933; Ph.D., 1936. Bell Telephone Laboratories, 1936–. Dr. Pierce has been engaged in the study of vacuum tubes.

Claude E. Shannon, B.S., in Electrical Engineering, University of Michigan, 1936; S.M. in Electrical Engineering and Ph.D. in Mathematics, M.I.T., 1940. National Research Fellow, 1940. Bell Telephone Laboratories, 1941–. Dr. Shannon has been engaged in mathematical research principally in the use of Boolean Algebra in switching, the theory of communication, and cryptography.

George C. Southworth, B.S., Grove City College, 1914; Sc.D. (Hon.), 1931; Ph.D., Yale University, 1923. Assistant Physicist, Bureau of Standards, 1917–18; Instructor, Yale University, 1918–23. Editorial staff of The Bell System Technical Journal, American Telephone and Telegraph Company, 1923–24; Department of Development and Research, 1924–34; Research Department, Bell Telephone Laboratories, 1934–. Dr. Southworth's work in the Bell System has been concerned chiefly with the development of the waveguide as a practical medium of transmission. He is the author of numerous papers relating to a diversity of subjects such as ultra-short waves, short-wave radio propagation, earth currents, the transmission of microwaves along hollow metal pipes and dielectric wires and microwave radiation from the sun.

VOLUME XXIX OCTOBER, 1950 NO. 4

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

50¢ per copy

$1.50 per Year

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the

American Telephone and Telegraph Company

195 Broadway, New York 7, N. Y.

SUBSCRIPTIONS

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

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XXIX *October, 1950* *No. 4*

Theory of Relation between Hole Concentration and Characteristics of Germanium Point Contacts

By J. BARDEEN

(*Manuscript Received Apr. 7, 1950*)

The theory of the relation between the current-voltage characteristic of a metal-point contact to *n*-type germanium and the concentration of holes in the vicinity of the contact is discussed. It is supposed that the hole concentration has been changed from the value corresponding to thermal equilibrium by hole injection from a neighboring contact (as in the transistor), by absorption of light or by application of a magnetic field (Suhl effect). The method of calculation is based on treating separately the characteristics of the barrier layer of the contact and the flow of holes in the body of the germanium. A linear relation between the low-voltage conductance of the contact and the hole concentration is derived and compared with data of Pearson and Suhl. Under conditions of no current flow the contact floats at a potential which bears a simple relation, previously found empirically, with the conductance. When a large reverse voltage is applied the current flow is linearly related to the hole concentration, as has been shown empirically by Haynes. The intrinsic current multiplication factor, α, of the contact can be derived from a knowledge of this relation.

I. Introduction

IN DISCUSSIONS of the theory of rectification at metal-semiconductor contacts, it is usually assumed that only one type of current carrier is involved: conduction electrons in *n*-type material or holes in *p*-type material.[1] In the case of metal-point contacts to high-purity *n*-type germanium, such as is used in transistors and high-back-voltage varistors, it is necessary to consider flow by both electrons and holes. A large part of the current in the direction of easy flow (metal point positive) consists of holes which flow into the *n*-type germanium and increase the conductivity of the material in the vicinity of the contact.[2,3] The conductivity is increased not only by the presence of the added holes but also by the additional conduction electrons which flow in to balance the positive space charge of the holes. There is a small concentration of holes normally present in the germanium under equilibrium conditions with no

[1] For a discussion of the nature of current flow in semi-conductors see the "Editorial Note" in *Bell Sys. Tech. Jour.* 28, 335 (1949).

[2] J. Bardeen and W. H. Brattain, *Bell Sys. Tech. Jour.* 28, 239 (1949).

[3] W. Shockley, G. L. Pearson and J. R. Haynes, *Bell Sys. Tech. Jour.* 28, 344 (1949).

current flow. When the contact is biased in the reverse (negative) direction, these holes tend to flow toward the contact and contribute to the current. The hole current is increased if the concentration of holes in the germanium is enhanced by injection from a neighboring contact or by creation of electron-hole pairs by light absorption.

Much has been learned about the effect of an added hole concentration on the current voltage characteristics of contacts from studies with germanium filaments. Part of this work is summarized in a recent article of W. Shockley, G. L. Pearson and J. R. Haynes.[3] These authors have investigated the way the low-voltage conductance of a point contact to a filament of *n*-type germanium varies with the concentration of holes in the filament and have shown that there is a linear relation between conductance and hole concentration. They have shown that the current to a contact biased with a large voltage in the reverse direction varies linearly with hole concentration. Suhl and Shockley[4] have shown that by applying a large transverse magnetic field along with a large current flow holes may be swept to one side of the filament. Changes in hole concentration produced in this way are detected by measuring changes in the conductance of a point contact.

Shockley[5] has suggested that the floating potential measured by a contact made to a semiconductor in which the concentration of carriers is not in thermal equilibrium may depend on the nature of the contact and differ from the potential in the interior. Pearson[6] has investigated this effect for point contacts on germanium filaments, and has shown that the floating potential is related to the conductance of the contact. This effect provides an explanation for anomalous values of floating potentials measured by Shockley[6] and by W. H. Brattain.[6] They found that potentials measured on a germanium surface in the vicinity of an emitter point biased in the forward direction may be considerably higher than expected from the conductivity of the material.

The purpose of the present paper is to develop the theory of these relations. We are particularly interested in effects produced by changes in hole concentration in *n*-type germanium resulting from hole injection or photoelectric effects. The equations developed also apply to injected electrons in *p*-type semiconductors with appropriate changes in signs of carriers and bias voltages. The methods of analysis used are similar to those which have been employed by Brattain and the author in a discussion of the forward current in germanium point contacts[2].

[4] H. Suhl and W. Shockley, *Phys. Rev.* 74, 232 (1948).
[5] W. Shockley, *Bell Sys. Tech. Jour.* 28, 435 (1949), p. 468.
[6] Unpublished.

The problem may be divided into two parts, which can be treated separately:

(a) The first deals with the current-voltage characteristics of the space charge region of the rectifying contact. The current flowing across the contact is expressed as the sum of the current which would flow if the hole concentration in the interior were normal and the current which results from the added hole concentration.

(b) The second is concerned with the current flow in the semiconductor outside the space charge region. In general, both diffusion and conduction

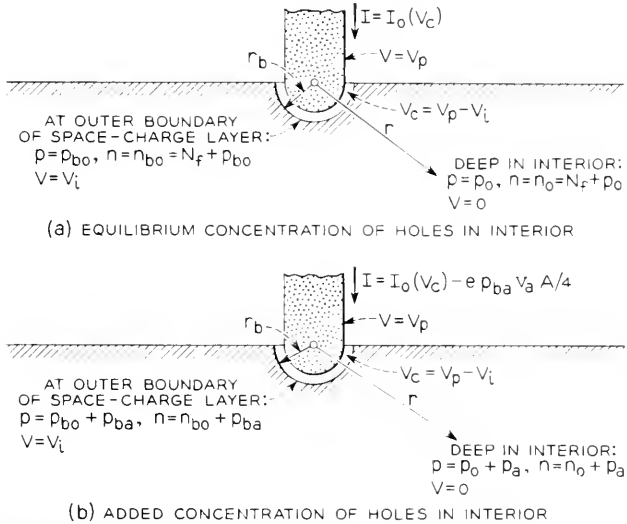

Fig. 1.—Model and notation used for calculation of current flow in low-voltage case.

are important in determining the flow of carriers, although, depending on conditions, one may be much more important than the other. In case the applied voltage and current flow are small, holes in an *n*-type semiconductor move mainly by diffusion. This situation applies to the problems discussed in the first part of the memorandum. In Section IV we discuss the opposite limiting case of large voltages in which the electron current flowing is so large that the hole current is determined by the electric field and diffusion is unimportant.

The model which is used to investigate the low-voltage case is illustrated in Fig. 1. For purposes of mathematical convenience, the contact is represented as a hemisphere extending into the germanium. Recombination, both at the surface of the semiconductor and in the interior, is

assumed to be negligible so that the lines of current flow are radial. The spherical symmetry of the resulting problem simplifies the mathematics. A calculation is given in an Appendix for a model in which the contact is a circular disk and recombination takes place at the surface. The latter does not give results which are significantly different from the simplified model.

Figure 1(a) applies to the case in which the hole concentration deep in the interior has its normal or thermal equilibrium value, p_0. The subscript zero is used to denote values which pertain to this situation. Of a voltage V_P applied to the contact, a part V_c occurs across the space-charge barrier layer of the contact and a part V_i occurs in the body of the semiconductor. Thus V_P represents the voltage of the contact and V_i the voltage in the semiconductor just outside the barrier layer, both measured relative to a point deep in the interior. It should be noted that V_P does *not* include the normal potential drop which occurs across the barrier layer under equilibrium conditions with no voltage applied. In the examples with which we shall deal in the present memorandum, the spreading resistance is small compared with the contact resistance, so that V_i is small compared with V_P. Obviously,

$$V_P = V_c + V_i. \tag{1}$$

When a current is flowing to the contact the hole concentration, p_{b0}, measured just outside of the barrier layer, differs from the concentration deep in the interior, p_0. It is the concentration gradient resulting from the difference between p_{b0} and p_0 which produces a flow of holes from the interior to the contact. In the forward direction, p_{b0} is larger than p_0; in the reverse direction, p_{b0} is less than p_0.

The total current, $I_0(V_c)$, flowing across the contact includes both electron and hole currents. It will not be necessary to distinguish between these two contributions to the normal current flow across the barrier layer in the subsequent analysis.

Figure 1(b) applies to the case in which the hole concentration deep in the interior has been increased to $p_0 + p_a$ by adding a concentration p_a to the normal concentration, p_0. The concentration just outside the barrier layer is increased to $p_{b0} + p_{ba}$. In addition to the normal current, $I_0(V_c)$, flowing across the contact, there is an additional current of holes resulting from the added hole concentration, p_{ba}, at the barrier.

The magnitude of this added hole current is determined in the following way. It is assumed that all holes which enter the barrier region are drawn into the contact by the field existing there. The number of holes

entering the barrier region per second is given by the following expression from kinetic theory:

$$p_b v_a A/4, \tag{2}$$

where v_a is the average thermal velocity, $2(2kT/\pi m)^{1/2}$, of a hole and A is the contact area. This expression gives the average number of particles which cross an area A from one side per second in a gas with concentration p_b. It follows that the current due to the added holes is:

$$I_{pa} = -ep_{ba}v_a A/4. \tag{3}$$

Since, by convention, a current flowing into the semiconductor is positive, a current of holes flowing from the interior to the contact is negative.

The diffusion current resulting from the added holes depends on the difference between p_{ba} and p_a. We shall show in Section III that when p_a is small compared with the normal electron concentration,

$$I_{pa} = 2\pi r_b kT\mu_p(p_{ba} - p_a), \tag{4}$$

where r_b is the radial distance to the outer boundary of the barrier layer and μ_p is the hole mobility. The value of p_{ba} is found by equating (4) and (3), i.e., the added current flowing from the interior to the barrier layer and the current flowing across the barrier layer. This gives

$$p_{ba}/p_a = a/(1 + a), \tag{5}$$

where a, defined by

$$a = 4(kT/er_b)\mu_p/v_a, \tag{6}$$

is the ratio of the velocity acquired by a hole in a field $4kT/er_b$ to thermal velocity. This ratio is generally a small number so that the a in the denominator of (5) can be neglected in comparison with unity. Equation (3) then becomes:

$$I_{pa} = -eap_a v_a A/4 = -p_a kT\mu_p A/r_b. \tag{7}$$

If p_a is not assumed small, a similar procedure may be used but the expressions for I_{pa} in terms of p_a are more complicated than (4) and (7)

It is possible that the added hole current, I_{pa}, will affect the contact in such a way as to change the normal current flowing. If there is such a change, one might expect it to be proportional to I_{pa} as long as I_{pa} is sufficiently small. The total current flow may then be expressed in terms of an "intrinsic α" for the contact as follows:

$$I = I_0(V_c) - \alpha I_{pa}(p_a). \tag{8}$$

There is no good theoretical reason to expect that α is different from unity for small current flow in normal contacts unless trapping is important.

Equation (8) is used as the basis for the analysis of the low-voltage data. One important consequence of the equation is that if p_a is different from zero, there is a voltage drop across the barrier layer even though no net current flows to the point. The presence of the added holes in the interior produces a floating potential on the point. The magnitude of this floating potential, V_{cf}, is obtained by setting $I = 0$ in Eq. (8) and finding the value of V_c which solves the equation. This potential can be observed on a voltmeter and is analogous to a photovoltage.

Associated with the floating potential is a change in conductance of the contact. The conductance near $I = 0$, given by

$$G = (dI\ dV_c)_{V_c=V_{cf}} = (dI_0/dV_c)_{V_c=V_{cf}}, \tag{9}$$

is just the conductance for normal hole concentration in the interior at an applied voltage equal to V_{cf}. In setting the conductance equal to the derivative of I with respect to V_c, we have neglected the difference, V_i, between V_c, the voltage drop across the barrier, and V_p, the total drop from the contact to the interior. This corresponds to neglecting the spreading resistance in comparison with the barrier resistance.

Equation (8) may be used to relate the floating potential with change of conductance of the contact. The appropriate equations, together with applications to data of Pearson and of Brattain, are given in Section II. In Section III we derive Eq. (4) which relates the added hole current with the added hole concentration in the interior. This relation is used to show that the point conductance G varies linearly with the added hole concentration, p_a. The theoretical expression for conductance is compared with data of Pearson and of Suhl.

In section IV we discuss the dependence of the current-voltage characteristic at large reverse voltages on hole concentration. Under these conditions it is the electric field rather than diffusion which produces the hole current in the body of the germanium. The electron and hole currents are then in the ratio of the electron to hole conductivity. With introduction of an "intrinsic α" for the contact, a simple relation is derived for the dependence of current on hole concentration for fixed voltage on the point. This relation is used to determine α for several point contacts from some data of J. R. Haynes.

II. Floating Potential of Point Contact

In order to get analytic expressions for the floating potential and admittance, it is necessary to make some assumption about the normal cur-

rent-voltage characteristic, $I_0(V_c)$. It is found empirically[7] that as long as V_c is not too large (a few tenths of a volt for a point contact on n-type germanium), it is a good approximation to take:

$$I_0 (V_c) = I_c (\exp(\beta e V_c / kT) - 1), \tag{10}$$

where I_c is a constant for a given contact. Except for the factor β, this is of the form to be expected from the diode theory of rectification. The empirical value of β is usually less than the theoretical value of unity in actual contacts.

If (10) is inserted into (8), the following equation is obtained for the current when there is an added concentration of carriers, p_a, in the interior:

$$I = I_c (\exp(\beta e V_c / kT) - 1) - \alpha I_{pa}. \tag{11}$$

Setting $I = 0$ and solving the resulting equation for the floating potential, $V_c = V_{cf}$, we find:

$$V_{cf} = (kT / e\beta) \log [1 + \alpha(I_{pa} / I_c)]. \tag{12}$$

The floating potential may be simply related to the conductance corresponding to small current flow. Using Eqs. (9) and (11), we find:

$$G = (dI_0/dV_c)_{V_c=V_{cf}} = (\beta e I_c / kT) \exp (\beta e V_{cf} / kT). \tag{13}$$

Since the normal low-voltage conductance is just

$$G_0 = \beta e I_c / kT, \tag{14}$$

we have

$$G = G_0 \exp (\beta e V_{cf} / kT). \tag{15}$$

By using (12), G can be expressed in terms of p_a. This relation is given and compared with experiment in Section III. Equation (15) may be solved for the floating potential:

$$V_{cf} = (kT / e\beta) \log (G / G_0). \tag{16}$$

It should be noted that (16) does not involve p_a directly. Thus it is possible to determine V_{cf} from a measurement of the change in conductance without direct knowledge of the added hole concentration. It holds for large as well as small p_a.

The logarithmic relation (16) between floating potential and conductance has been demonstrated by an experiment of Pearson. The experi-

[7] See H. C. Torrey and C. A. Whitmer, "Crystal Rectifiers", McGraw-Hill Company, New York, N. Y., (1949), p. 372–377.

mental arrangement is illustrated in Fig. 2. Holes are injected into a germanium filament by an emitter point and the circuit is closed by allowing the current to flow to the large electrode at the left end. The right end of the filament is left floating. Some of the injected holes diffuse

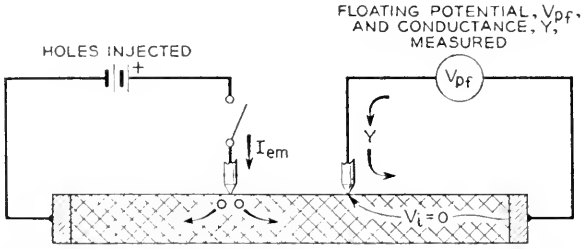

Fig. 2.—Schematic diagram of experiment of G. L. Pearson to investigate relation between floating potential and impedance of point contact.

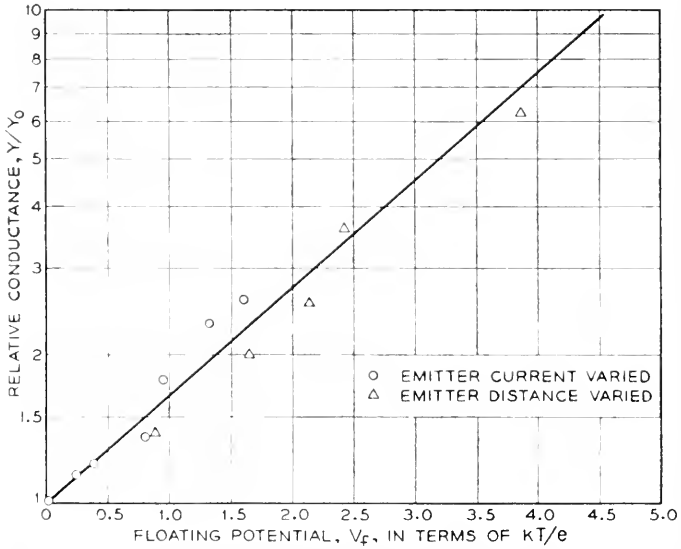

Fig. 3.—The relationship of admittance ratio to potential, measured at a point on a germanium filament into which holes are emitted, with no current flow, from G. L. Pearson's data of September 21, 1948.

down the filament and increase the local concentration in the neighborhood of the probe point. This concentration can be varied by changing the emitter current and also by changing the distance between emitter and probe. Both the floating potential and the conductance between the probe point and the large electrode on the right end were measured. Under the conditions of this experiment, the potential drop in the in-

terior of the floating end of the filament is small. The small drop which does exist results from the difference in mobility between electrons and holes. Almost all of the potential difference between the probe and the right end is the floating potential, V_{cf}, across the barrier layer of the probe point.

Pearson's data are plotted in Fig. 3. The data can be fitted by an equation of the form (16) with $\beta = 0.5$.

The difference in potential between a floating point contact and the interior which exists under non-equilibrium conditions explains anomalously high values of probe potential which were sometimes observed by Shockley and by Brattain in the vicinity of an emitter point operating in the forward direction. As an example of a case in which the effect is

TABLE I

Measurements of probe potential, V_{pf}, at a contact on an etched germanium surface .005 cm from a second contact carrying a current I. The conductance of the probe point is G_P. The voltage drop across the probe contact, $V_{pf} - V_i$, at zero current is calculated from $V_{pf} - V_i = 2.5(kT/e) \log (G_p/G_0)$. Data from W. H. Brattain.

I amps	V_{pf} volts	V_{pf}/I ohms	G_p mhos	G_p/G_0	$\log (G_p/G_0)$	$V_{pf} - V_i = .062 \log (Y_p/Y_e)$	V_i volts	V_i/I ohms
2.0×10^{-3}	0.189	94	8.3×10^{-4}	6.9	1.93	0.120	0.069	35
1.0	0.141	141	5.0	4.2	1.435	0.090	0.051	51
0.5	0.096	190	3.3	2.8	1.030	0.064	0.032	64
0.2	0.052	260	2.2	1.8	0.588	0.037	0.015	75
0.1	0.030	300	1.7	1.4	0.336	0.021	0.009	90
−0.1	−0.0096	96	1.2	1.0				
−0.2	−0.0186	93	1.2	1.0				
−0.5	−0.044	88	1.25	—				
−1.0	−0.10	100	1.35	—				

large, some data of Brattain are given in Table I for the experimental arrangement of Fig. 4. Two point contacts were placed about .005 cm apart on the upper face of a germanium block. The surface was ground and etched in the usual way. A large-area, low-resistance contact was placed on the base. The potential, V_P, of one point, used as a probe, was measured as a function of the current flowing in the second point. In this case, the potential on the probe point is produced in part by the V_{cf} term and in part by a potential, V_i, in the interior which comes from the IR drop of the current flowing from the emitter point to the base electrode. Reasonable values are obtained for V_i from measurements of V_P if a correction for V_{cf} is properly made.

The first column of Table I gives the current and the second column the probe potential, V_P, measured relative to the base. The third column gives values of V_P/I. In the reverse direction (negative currents) V_P/I

is approximately constant at a little less than 100. Values of V_p/I in the forward direction are much larger, starting at 300 for $I = 0.1$ ma and decreasing to 94 at $I = 2$ ma. If anything, one would expect a decrease rather than an increase in V_p/I in the forward direction as injection of holes lowers the resistivity of the germanium in the vicinity of the point. We shall show that V_i/I actually does decrease and that the anomalously high values of V_p/I in the forward direction result from the drop, V_{ef},

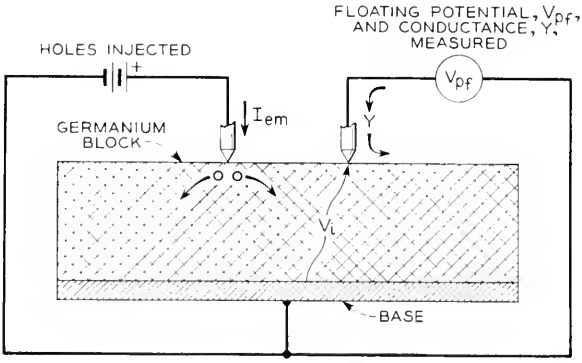

Fig. 4.—Schematic diagram of experiment of W. H. Brattain for measuring floating potential and admittance at point near emitter.

across the barrier layer between the contact point and the body of the germanium. Thus,

$$V_i = V_p - V_{ef}. \tag{17}$$

Values of V_{ef} can be estimated from the change in conductance corresponding to small currents in the probe point. The conductance increases with increasing forward emitter current. Values of V_{ef}, calculated from

$$V_{ef} = 2.5\ (kT/e) \log\ (G_p/G_0), \tag{18}$$

are given in column 6. The value 2.5, chosen empirically to give reasonable values of V_i, is not far from the value 2.0 required to fit Pearson's data in Fig. 1. Values of V_i obtained from Eq. (17) are given in column 7. The ratios V_i/I given in column 8 are reasonable. The decrease in V_i/I with increasing forward current is caused by a decrease in the resistivity of the germanium resulting from hole injection.

In another case, in which no such anomaly was observed in the forward direction, it was found that V_{ef}, calculated from the change in conductance, was small compared with V_p.

There have as yet been no measurements which permit a comparison of the values of β required to correlate probe potential and conductance

with values of β obtained directly from the current-voltage characteristic of the probe. Such a comparison would provide a valuable test of the theory.

III. Low Voltage Conductance of Point Contacts

In this section we calculate the hole current flowing in the body of the germanium from diffusion and find an expression relating change of conductance with added hole concentration. The results shall be applied to data of Pearson and of Suhl. We need to derive Eq. (4) which gives the hole current in terms of the added hole concentrations, p_{ba}, measured just outside the barrier layer, and p_a, measured deep in the interior.

The model which is used for the calculation is illustrated in Fig. 1. The diffusion equation for hole flow is to be solved subject to the boundary conditions that $p = p_b$ just outside the barrier layer and $p = p_i$ at large distances from the contact in the interior. It is assumed that the total current flow is zero or small.

We shall first derive the more general equations[8] which include flow by the electric field as well as by diffusion in order to show the conditions under which the electric field can be neglected. In the body of the semiconductor, conditions of electric neutrality require that the electron concentration, n, be given by:

$$n = N_f + p, \tag{19}$$

where N_f, the net concentration of fixed charge, is the difference between the concentrations of donor and acceptor ions. We shall assume that N_f is constant so that

$$\text{grad } n = \text{grad } p. \tag{20}$$

The general equations for electron and hole current densities, i_n and i_p, are:

$$i_n = \mu_n (enF + kT \text{ grad } n) \tag{21}$$

$$i_p = \mu_p (epF - kT \text{ grad } p), \tag{22}$$

where F is the electric field strength. By using (19) and (20), and setting $\mu_n = \beta\mu_p$, we can express i_n in the form:

$$i_n = b\mu_p (e(N_f + p) F + kT \text{ grad } p). \tag{23}$$

The magnitude of F for zero net current,

$$i = i_p + i_n = 0, \tag{24}$$

[8] A discussion of the equations of flow is given in the article by W. van Roosbroeck in this issue of the *Bell System Technical Journal*.

can be obtained by adding (22) and (23) and equating the result to zero. This gives:

$$\frac{eF}{kT} = -\frac{b-1}{N_f b + p(b+1)} \text{ grad } p. \tag{25}$$

The field vanishes for $b = 1$, corresponding to equal mobilities for holes and electrons. For b greater than unity and for equal concentration gradients of holes and electrons, the diffusion current of electrons is larger than that of holes. The field is such as to equate these currents by increasing the flow of holes and decreasing the flow of electrons.

If (25) is substituted into (22), the following equation is obtained for i_p:

$$i_p = -kT\mu_p \left[\frac{(b-1)p}{N_f b + p(b+1)} + 1 \right] \text{grad } p. \tag{26}$$

If recombination is neglected, the hole current is conserved and

$$\text{div } i_p = 0. \tag{27}$$

Using this relation, an equation of the Laplace type can be obtained for p which may be integrated subject to the appropriate boundary conditions. This derivation is given in Appendix B. The results do not differ significantly from those obtained below for p assumed small.

Rather than continue with the general case, we shall at this point assume that $p \ll N_f$ so that the first term in the parenthesis of Eq. (26) is negligible in comparison with unity. This amounts to setting $F = 0$ in Eq. (3) and assuming that the holes move entirely by diffusion. This is a very good approximation in most cases of practical interest and is valid for small i as well as for $i = 0$. We then have

$$i_p = -kT\mu_p \text{ grad } p. \tag{28}$$

The condition div $i_p = 0$ gives Laplace's equation for p:

$$\nabla^2 p = 0. \tag{29}$$

Equation (29) is to be solved subject to the appropriate boundary conditions. For the model illustrated in Fig. 1 we can assume that p depends only on the radial distance r and that

$$p = p_b \text{ at } r = r_b, \tag{30}$$

$$p = p_i \text{ at } r = \infty. \tag{31}$$

The solution of (29) which satisfies (31) is:

$$p = p_i + (I_p/2\pi kT\mu_p r), \tag{32}$$

in which I_p is the total hole current. The boundary condition (30) gives the relation between I_p and p_b:

$$p_b = p_i + (I_p/2\pi kT\mu_p r_b). \tag{33}$$

Since the equations are linear, an equation of the form (33) applies to the hole current due to the added holes as well as to the entire hole current. For the former we have:

$$p_{ba} = p_a + (I_{pa}/2\pi kT\mu_p r_b), \tag{34}$$

which is equivalent to Eq. (4).

In the derivation of Eq. (34) we have neglected recombination at the surface as well as in the interior. In the Appendix we give a solution for a contact in the form of a circular disk and assume that recombination takes place at the surface. The hole concentration then satisfies Laplace's equation subject to more complicated boundary conditions at the surface. The results are not significantly different from those of the simplified model.[9]

Equation (34), or rather its equivalent, Eq. (4), was used in the derivation of Eq. (12) for the floating potential, V_{ef}. If this value for V_{ef} is inserted into Eq. (15), an equation relating the conductance directly with the added hole concentration is obtained:

$$G = G_0 + (\alpha e^2 a v_a A \beta p_a/4kT). \tag{35}$$

This expression may be simplified by substituting for a from Eq. (6):

$$G = G_0 + \alpha\beta\mu_p e A p_a/r_b. \tag{36}$$

By using the expression for the normal conductivity:

$$\sigma_0 = b\mu_p e n_0, \tag{37}$$

the conductance can be given in the form:

$$G = G_0 + (\alpha\beta\sigma_0 A/br_b)(p_a/n_0). \tag{38}$$

If σ_0 is in practical units (mhos/cm), G is in mhos.

We shall compare (38), which gives a linear variation between G and p_a, with experimental data of Pearson[10] and of Suhl. The arrangement used

[9] In the applications, these equations are applied to situations in which the contact is on a germanium filament and there is a flow of current along the length of the filament in addition to the flow to the contact. A question may arise as to whether it is justified to neglect the filament current when discussing flow to the contact. There is no difficulty as long as p_a/n_0 is small compared with unity because the equations are then linear and the solution giving the flow to the contact can be superimposed on the solution giving the flow along the length of the filament. The neglect of the filament current cannot be rigorously justified in case p_a/n_0 is large, as is assumed in the calculations of Appendix B. It is not believed, however, that the exact treatment would yield results which are significantly different.

[10] See reference 3, p. 356 and Fig. 6.

by Pearson is shown in Fig. 5. Two probe points were placed about .009 cm apart near one end of a germanium filament. The concentration

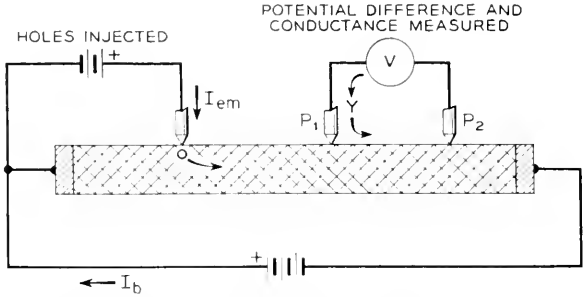

Fig. 5.—Experimental arrangement used by G. L. Pearson to investigate relation between admittance and hole concentration in germanium filament.

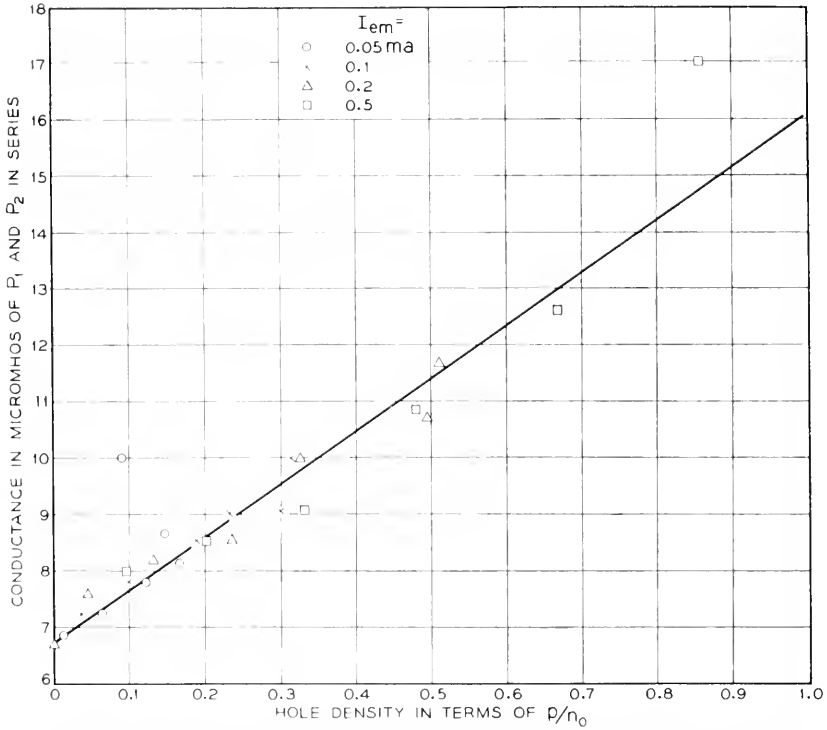

Fig. 6. The relationship between point admittance and relative hole concentration, for a germanium filament from G. L. Pearson's data of September 28, 1948.

of holes was varied by current from an emitter point near the opposite end of the filament. There was an additional current flowing between

electrodes at the two ends so that the field pulling the holes along the filament could be varied. The concentration of holes was determined from the change in resistivity of that segment of the filament between the two probes. Measurements of admittance were made by passing a small current between the two probes connected in series. The area of the filament is about 1.6×10^{-4} cm^2 and the normal resistance between the probes about 1800 ohms. The normal conductivity is thus

$$\sigma_0 = .009/(1800 \times 1.6 \times 10^{-4}) = 0.03 \text{ (ohm cm)}^{-1}. \tag{39}$$

As shown in Fig. 6, Pearson finds a linear relation between G and p_a. The line best fitting Pearson's data is

$$G = G_0 + (8 \times 10^{-6})\ (p_a/n_0) \text{ (mhos)}. \tag{40}$$

The theoretical value of the coefficient may be obtained from Eq. (38). Taking

$$\begin{aligned} &\alpha = 1, \quad \beta = 0.5, \qquad \sigma_0 = 0.03 \\ &b = 2.0, \ A = 10^{-6} \text{ cm}^2, \ r = 5 \times 10^{-4} \text{ cm}, \end{aligned} \tag{41}$$

we get

$$\alpha\beta\sigma_0 A/b\ r_b = 15 \times 10^{-6} \text{ mhos}. \tag{42}$$

Pearson's data, represented by (40), apply to the conductance of two point contacts in series, and the conductance of each one may be about twice that given by (40). Thus the theoretical value is in good agreement with the observed. There is no indication that α differs from unity at low voltage.

Suhl varied the concentration of holes in the vicinity of probe points by application of a transverse magnetic field as well as by injection from an emitter point. The experiment is illustrated in Fig. 7. He used a filament with a cross-section of about $.025 \times .025$ cm. Four probe points were placed along the length of the filament at intervals of about .04 cm. A total current of 4 ma flowed in the filament.

In one experiment, none of this current was injected, so that the concentration of holes was normal in the absence of the magnetic field. Measurements were made of the floating potentials and of the conductances of the probe points. Then a transverse magnetic field was applied and the conductances measured again. We are interested here only in the case of a large field (30,000 gauss) in such a direction as to sweep the holes to the opposite side of the filament. Suhl believes that under these conditions the concentration of holes near the probe points is practically zero. The difference between the conductances with and without the field

then gives the contribution to the conductance from the normal concentration of holes.

In a second experiment 1 ma of the current of 4 ma flowing in the filament was injected from an emitter point near one end of the filament. From the probe potentials, estimates have been made of the change in

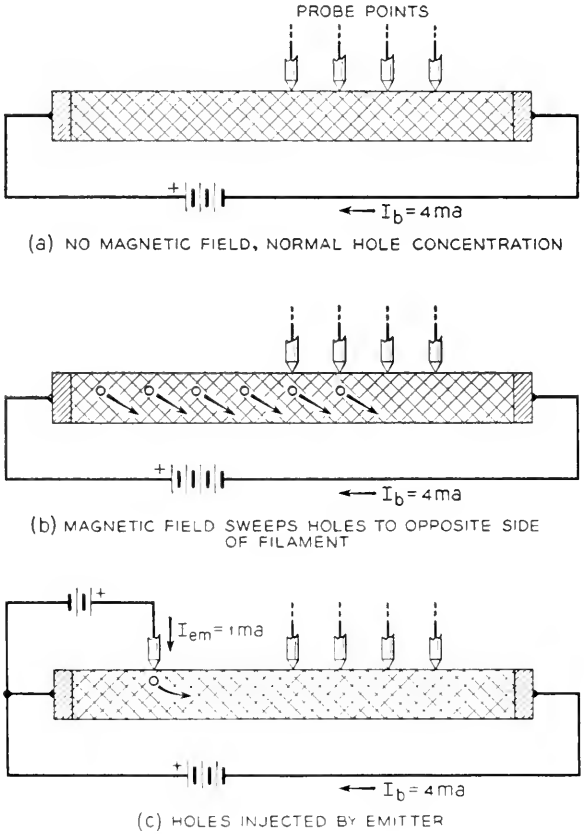

Fig. 7.—Schematic diagram of experiment of H. Suhl to investigate relation between hole concentration and impedance of point contacts.

resistivity and thus of the added hole concentration at the different probe points. Changes in hole concentration from injection have been correlated with changes in admittance of the probe points.

The filament with dimensions .025 × .025 × 0.4 cm has a resistance of 4,600 ohms. The normal resistivity, ρ_0, is then about 7.2 ohm cm. Since the concentration of electrons corresponding to 1.0 ohm cm is about

1.8×10^{15}, the concentration corresponding to a resistivity of 7.2 ohm cm is[11]:

$$n_0 = 1.8 \times 10^{15}/7.2 = 2.5 \times 10^{14}/\text{cm}^3. \tag{43}$$

The product of the equilibrium concentrations of electrons and holes is about 4×10^{26} in germanium at room temperature[12]. Thus, for this sample,

$$p_0 = 4 \times 10^{26}/2.5 \times 10^{14} = 1.5 \times 10^{12}/\text{cm}^3. \tag{44}$$

If there is an added concentration of holes, p_a, resulting from injection, the added conductivity is:

$$\sigma_a = (1 + b)\, e\mu_h p_a = 8.4 \times 10^{-16}\, p_a\,. \tag{45}$$

The resistivity is changed to:

$$\rho = \rho_0\sigma_0/(\sigma_a + \sigma_0) \simeq \rho_0\,(1 - \sigma_a\rho_0), \tag{46}$$

the approximate expression holding if the relative change is small. The resistance per unit length of filament is:

$$R = 1.15 \times 10^4\,(1 - \sigma_a\rho_0). \tag{47}$$

The change in voltage gradient, $dV/dx = RI$, resulting from hole injection is, for a current of 4×10^{-3} amps,

$$\Delta(dV/dx) = d(\Delta V)/dx = -46\rho_0\sigma_a\,. \tag{48}$$

Suhl measured the change in probe potential, ΔV, which resulted when 1 ma of the total current of 4 ma was injected from the emitter instead of having the entire 4 ma flowing between the ends of the filament. His values of ΔV for the four probe points are given in Table II. We have made a plot of these as a function of position and have estimated the gradients at each of the four probe positions. Using these values we have calculated σ_a from Eq. (48) and the corresponding injected hole concentration from Eq. (45). These are given in the last column of the table.

Suhl's measurements of conductances, G, of the probe points are given in Table III. Also given are differences, ΔG, from the normal values with no magnetic field and no injection and also these differences multiplied by n_0/p_a. Values of p_a for the case of hole injection were obtained from Table II. Values of $\Delta G(n_0/p_a)$ are to be compared with the theoretical value,

$$\Delta G(n_0/p_a) = \alpha\beta\sigma_0 A/cr_b\,, \tag{49}$$

[11] These values are based on taking $\mu_n = 3500$ cm²/volt sec and $\mu_p = 1700$ cm²/volt sec, as measured by J. R. Haynes. They correspond to room temperature (295°K).

[12] This value is obtained from an intrinsic resistivity of about 60 ohm cm for *Ge* at room temperature and the mobility values in reference 11.

from Eq. (38). Taking $\alpha = 1, \beta = 0.5, \sigma_0 = 0.14, b = 1.5, A = 10^{-6}$ and $r_b = 5 \times 10^{-4}$, we get

$$G(n_0/p_a) \sim 100 \text{ micromhos.} \tag{50}$$

This value is of the same order as the values obtained from Suhl's data listed in Table III. There is a large scatter in the latter and the values are

TABLE II

Calculation of hole concentrations from probe potential measurements. ΔV measures potential difference resulting from hole injection of 1 ma when total current is kept at 4 ma; data from H. Suhl.

Point No.	Relative Position (cm)	ΔV (volts)	$\frac{d\Delta V}{dx}$ (volts/cm)	$\rho_0\sigma_a$	σ_a (mhos)	p_a (cm^{-3})
#6	0	−.04	−0.6	.013	.0018	2.2×10^{12}
#5	.044	−.073	−1.10	.024	.0033	4.0
#4	.084	−.13	−1.8	.039	.0054	6.5
#3	.12	−.21	−2.5	.055	.0077	9.0

TABLE III

Changes in conductance resulting from application of magnetic field and from hole injection. Units are micromhos. Data from H. Suhl.

Point	No Field	With—30,000 gauss field			With hole injection		
	G	G	ΔG	$\Delta G \frac{n_0}{(-p_0)}$	G	ΔG	$\Delta G \frac{n_0}{p_a}$
#6	17.2	16.4	−0.8	130	22.5	7.8	880
#5	6.55	4.35	−2.2	365	7.0	0.45	28
#4	3.7	3.2	−0.5	80	5.1	1.4	54
#3	13.0	9.2	−3.8	630	19	6	165

not consistent. It has been suggested that the abnormal values may result from local sources of holes.

IV. Hole Flow for a Collector with Large Reverse Voltage

Haynes has shown that there is a linear relation between the current to a collector point operated in the reverse direction and the concentration of holes in the interior of a germanium filament. Under the conditions of his experiment, the current flowing to the collector point is small compared with the total current flowing down the filament, so that the collector current does not alter the concentrations very much. Holes are injected into the filament by an emitter point placed near one end, and the concentration is determined from the change in resistance of the filament in the neighborhood of the collector point.

Haynes' measurements may be fitted by an empirical equation of the following form:

$$I = I_0[1 + \gamma p_a/n_0)], \tag{51}$$

in which I_0 is the normal collector current flow for a given collector voltage, I is the collector current flowing for the same collector voltage when the hole concentration is increased by p_a, and n_0 is the normal electron concentration. Values of I_0 and γ for four different formed phosphor-bronze collector points are given in Table IV. The collector bias is -20 volts in each case. It can be seen that the variations in γ are much less than those in I_0. It will be shown below that γ is related to the intrinsic α of the point contact.

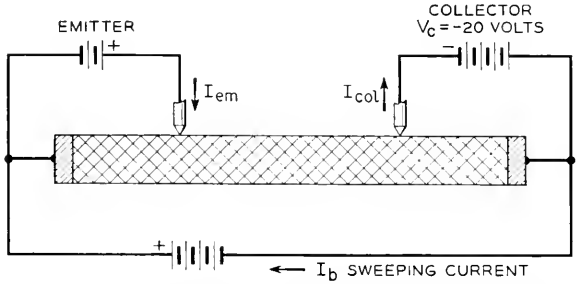

Fig. 8.—Experimental arrangement used by J. R. Haynes to determine relation between hole concentration and current to collector point biased with large voltage in reverse direction.

In Haynes' experiment, holes are attracted to the collector by the field produced by the electron current and diffusion plays a minor role. In contrast to the preceding examples, the terms involving the field F in Eqs. (21) and (22) are large and the diffusion terms represented by the concentration gradients are small. It follows from (21) and (22) that the ratio of electron to hole current density is then:

$$i_n/i_p = bn/p, \tag{52}$$

which is equal to the ratio of the electron and hole contributions to the conductivity. If n and p do not vary with position, the ratio is the same everywhere and equal to the ratio of total electron and hole currents, I_n and I_p:

$$I_n/I_p = i_n/i_p = bn/p. \tag{53}$$

The currents I_n and I_p can also be related to the intrinsic α for the contact by use of an equation of the form:

$$I = I_{n0} + \alpha I_p, \tag{54}$$

in which I_{n0} is the electron current for zero hole current. The electron current is:

$$I_n = I_{n0} + (\alpha - 1)I_p . \tag{55}$$

Thus we have

$$\frac{I_n}{I_p} = \frac{I_{n0} + (\alpha - 1)I_p}{I_p} = \frac{bn}{p} = \frac{b(N_f + p)}{p}. \tag{56}$$

This equation may be solved for I_p to give:

$$I_p = pI_{n0}/(bN_f + (\alpha - 1 - b)p). \tag{57}$$

The term $(\alpha - 1 - b)p$ is generally small compared with bN_f and may be neglected. We thus have approximately for p/N_f small and $N_f \simeq n_0$,

$$I = I_{n0} + \alpha I_p = I_{n0}[1 + (\alpha p/bn_0)]. \tag{58}$$

When expressed in terms of the normal current,

$$I_0 = I_{n0}[1 + (\alpha p_0/bn_0)], \tag{59}$$

the equation for I is of the form (51):

$$I = I_0 [1 + (\alpha p_a/bn_0)]. \tag{60}$$

From a comparison of (51) and (60) it can be seen that:

$$\gamma = \alpha/b \text{ or } \alpha = b\gamma. \tag{61}$$

Values of α determined from empirical values of γ for the four point contacts of Haynes are given in Table IV. The values are of a reasonable order of magnitude for formed collector points.

An estimate of the importance of diffusion can be obtained by comparing the hole current in Haynes' experiments with the hole current which would exist if the electron current were zero, so that holes move by diffusion alone. Equations (28) to (33) apply to the latter case. In addition to (33) we need an equation which expresses the hole current flowing into the contact in terms of the hole concentration, p_b, at the contact. If the reverse bias is large, no holes will flow out and the entire hole current is that from semiconductor to metal as given by an equation similar to (3):

$$I_p = -cp_b v_a A/4. \tag{62}$$

Substituting this value for I_p into equation (33) we get an equation which may be solved for p_b, to give:

$$p_b = ap_i(1 + a) \simeq ap_i, \tag{63}$$

with a given by Eq. (6). Using (63) for p_b, we get:

$$I_p = kT\mu_h p_i A/r_b = (kT\sigma_0 A/ebr_b)(p_i/n_0). \tag{64}$$

With $kT/e = .025$ volts, $\sigma_0 = bn_0e\mu_h = 0.2$ (ohm cm)$^{-1}$, $A = 10^{-6}$ cm^2 and $r_b = 5 \times 10^{-4}$ cm, we get for the diffusion current:

$$I_p = (5 \times 10^{-6})(p_i/n_0) \text{ amps.} \tag{65}$$

Comparing (65) with (57) we see that diffusion of holes will not be important if

$$I_{n0} \gg 5 \times 10^{-6} \text{ amps.} \tag{66}$$

This condition is satisfied in Haynes' experiments.

In the case of point contacts formed to have a high reverse resistance as diodes, I_0 may be of the order of 10^{-7} to 10^{-6} amps at room temperature. Diffusion of holes will then play a role, and the hole current will

TABLE IV

Relation between hole concentration and collector current from data of J. R. Haynes. Data represented by

$$I = I_0(1 + (\gamma p_a/n_0))$$

where I is current flowing to collector point biased at -20 volts and p_a/n_a is ratio of added hole concentration to the normal electron concentration.

Probe Point	I_0	$a = 2.1\gamma$
0	0.94	4.6
2	0.33	4.4
3	0.54	6.9
4	1.20	4.6

be larger than indicated by Eq. (53). As discussed in reference (4) there is still a question as to the importance of holes in the saturation current observed by Benzer in diodes with high reverse resistance. Experiments similar to those of Haynes would be valuable to determine the influence of hole concentration on reverse current.

ACKNOWLEDGMENT

The author is indebted to G. L. Pearson, J. R. Haynes, W. H. Brattain, and H. Suhl for use of the experimental data presented herein; to W. Shockley for a critical reading of the manuscript and a number of valuable suggestions, and to W. van Roosbroeck for aid with some of the analyses and for suggestions concerning the manuscript.

APPENDIX A

DIFFUSION OF HOLES WITH SURFACE RECOMBINATION

In the calculation of the diffusion of holes given in Section III of the text it was assumed that no recombination of electrons and holes oc-

curred. In the present calculation it is assumed that recombination occurs at the surface, but not in the volume. This is a good approximation for a point contact on germanium. It is further assumed that the hole concentration is sufficiently small so that Laplace's equation (29) may be used.

The model which we shall use is illustrated in Fig. 9. The contact is in the form of a circular disk of radius ρ on the surface of the semiconductor. Cylindrical coordinates, r, θ, z, are used, with the origin at the center of the disk and the positive direction of the z-axis running into the semiconductor. We calculate the flow due to the added holes, and shall use the symbol p without subscript to denote the added hole concentration.

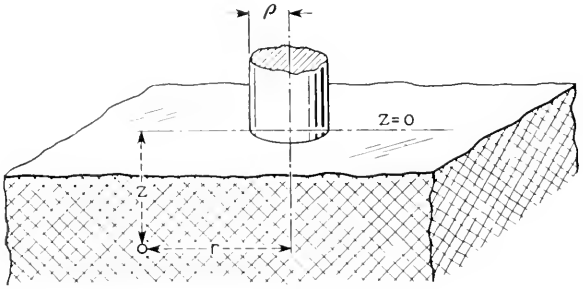

Fig. 9.—Coordinates used for calculation of hole flow to contact area in form of circular disk.

With recombination at the surface, it is necessary to have a gradient in the interior which brings the holes to the surface.

It is assumed that the rate of recombination at the surface is:

$$sp = \text{holes/cm}^2, \tag{1A}$$

where the factor s has the dimensions of a velocity and p is evaluated at the surface $z = 0$. According to measurements of Suhl and Shockley, s is about 1500 cm/sec for a germanium surface treated with the ordinary etch. The current flowing to the surface is:

$$(\mu_p kT/e)(\partial p/\partial z)_{z=0} \text{ holes/cm}^2. \tag{2A}$$

The boundary condition for p at the surface $z = 0$ outside of the contact area is obtained by equating (1A) and (2A). This gives:

$$\partial p/\partial z = \lambda p \text{ at } z = 0,\ r > \rho \tag{3A}$$

where

$$\lambda = se/\mu_p kT, \tag{4A}$$

has the dimensions of a length. For $s = 1500$ cm/sec and $\mu_p = 1700$ cm^2/volt sec, corresponding to germanium at room temperature, λ is about 35 cm^{-1}.

The boundary condition on the disk is similar to (3A) except that s is replaced by $v_a/4$ (cf. Eq. (3)). Thus for $r < \rho$,

$$\partial p/\partial z = \lambda_e p \qquad z = 0,\ r < \rho, \tag{5A}$$

where

$$\lambda_e = v_a e/4\mu_p kT. \tag{6A}$$

Evaluated for germanium at room temperature, λ_e is about 6×10^4.

In order to have a dependent variable which vanishes at infinity, we replace p by:

$$y = p_a - p + \lambda p_a z, \tag{7A}$$

so that $p \to p_a$ for $z = 0$ as $r \to \infty$. The variable y satisfies Laplace's equation subject to the boundary conditions:

$$\partial y/\partial z = \lambda y \qquad z = 0,\ r > \rho \tag{8A}$$

$$\partial y/\partial z = \lambda_e (y - p_a)\ z = 0,\ r < \rho \tag{9A}$$

$$y = 0 \qquad r, z \to \infty. \tag{10A}$$

An exact solution of the problem is difficult. We shall obtain an approximate solution which satisfies (8A) but not (9A) and which applies when

$$\lambda\rho \ll 1 \ll \lambda_e\rho. \tag{11A}$$

This approximation is valid for a germanium point contact, since, for $\rho \sim 10^{-3}$ cm,

$$\lambda\rho \sim .035,\ \lambda_e\rho \sim 60. \tag{12A}$$

We shall first discuss the limiting case for which $\lambda \to 0$ and $\lambda_e \to \infty$. The former implies neglect of surface recombination and the latter

$$y = p_a \text{ for } z = 0,\ r < \rho. \tag{13A}$$

The problem is the same as that of finding the potential due to a conducting circular disk. The solution of this problem, which is well known, is:

$$y = (2p_a/\pi) \int_0^\infty e^{-zt} J_0(rt) \frac{\sin \rho t}{t}\, dt. \tag{14A}$$

The current flowing to the disk is obtained from integrating:

$$i_z = kT\mu_p(\partial y/\partial z), \tag{15A}$$

over the area of the disk. This gives:

$$I_{pa} = -4\rho p_a kT\mu_p. \tag{16A}$$

The analogous expression for a hemispherical contact area of radius r_b, obtained from (7), is:

$$I_{pa} = -2\pi r_b p_a kT\mu_p. \tag{17A}$$

If a comparison is made on the basis of equal radii, (17A) is larger than (16A) by a factor of $\pi/2$. On the more reasonable basis of equal contact areas, (16A) is larger than (17A) by a factor of $4/\pi$.

An approximate solution which includes surface recombination can be obtained as follows. A solution of Laplace's equation which satisfies (8A) and (10A) is:

$$y = \frac{2y_0}{\pi}\int_0^\infty e^{-zt}J_0(rt)\frac{\sin \rho t}{t+\lambda}\,dt. \tag{18A}$$

That (18A) satisfies (8A) may be verified by direct substitution:

$$\left[-\frac{\partial y}{\partial z} + \lambda y\right]_{z=0} = \frac{2y_0}{\pi}\int_0^\infty J_0(rt)\sin \rho t\,dt = 0 \quad \text{for} \quad r > \rho. \tag{19A}$$

$$= (2y_0/\pi)(\rho^2 - r^2)^{-1/2} \quad \text{for} \quad r < \rho. \tag{20A}$$

Expression (18A) satisfies (9A) approximately if λ_c is large. Using (20A) and neglecting λ in comparison with λ_c, we have:

$$y = p_a - (2y_0/\pi\lambda_c)(\rho^2 - r^2)^{-1/2} \quad \text{for} \quad z = 0,\ r < \rho. \tag{21A}$$

Except for r almost equal to ρ, the second term on the right of (21A) is very small. It is not possible to obtain an explicit expression for y for $r < \rho$. For $z = 0$, $r = \rho$,

$$y = \frac{2y_0}{\pi}\int_0^\infty \frac{J_0(\rho t)\sin \rho t}{t+\lambda}\,dt = y_0 F(\lambda\rho). \tag{22A}$$

The integral, $F(\lambda\rho)$, can be evaluated from a more general integral in Watson's Bessel Functions, p. 433. We have:

$$F(k) = \frac{2}{\pi}\int_0^\infty \frac{J_0(x)\sin x\,dx}{x+k} = \cos k\,J_0(k) + \sin k\,\mathbf{I}_0(k). \tag{22B}$$

The factor multiplying y_0 is unity for $\lambda\rho = 0$, and decreases as $\lambda\rho$ increases. Since y is approximately equal to p_a, we have, approximately,

$$y_0 = p_a/F(\lambda\rho). \tag{23A}$$

The value of y can also be found for $r = 0$. For $z = 0$, $r = 0$, we have:

$$y = \frac{2y_0}{\pi} \int \frac{\sin \rho t \, dt}{t + \lambda} = \frac{2y_0}{\pi} G(\lambda\rho). \tag{24A}$$

The integral can be expressed in terms of integral sine and cosine functions:

$$G(k) = \frac{2}{\pi} \int_0^\infty \frac{\sin x \, dx}{x + \lambda} = \frac{2}{\pi} \left[-\cos k \left(\mathrm{Si}\ k - \frac{\pi}{2} \right) + \sin k \ \mathrm{Ci}\ k \right]. \tag{25A}$$

If k is not too large, $G(k)$ is nearly equal to $F(k)$, so that y is approximately constant over the area of the disk.

The total current flowing from the contact is found from integrating $kT\mu_p \, (\partial y/\partial z)$ over the disk:

$$I_{pa} = -kT\mu_p y_0 \int_0^\rho \int_0^\infty \frac{4rtJ_0(rt) \sin \rho t}{t + \lambda} \, dt \, dr \tag{26A}$$

$$= -4kT\mu_p y_0 \int_0^\infty \frac{\rho J_1(\rho t) \sin \rho t}{t + \lambda} \, dt. \tag{27A}$$

The integral can be evaluated with use of the general integral of Watson, to give:

$$I_{pa} = -4\rho kT\mu_p y_0 \, H(\lambda\rho), \tag{28A}$$

where

$$H(k) = \int_0^\infty \frac{J_1(x) \sin x \, dx}{x + k} = -\frac{\pi}{2} [\cos k \, J_1(k) + \sin k \, Y_1(k)]. \tag{29A}$$

Using (23A) for y_0, we have:

$$I_{pa} = -4\rho kT\mu_p p_a [H(\lambda\rho)/F(\lambda\rho)]. \tag{30A}$$

Except for the factor $H(\lambda\rho)/F(\lambda\rho)$, this expression for the current is identical with (16A). This factor, which gives the effect of recombination on the current, is plotted in Fig. 10. Recombination gives an increase in current flow, but the effect is small for the normal rate of surface recombination, which corresponds to $k = \lambda\rho \sim .035$.

APPENDIX B

Calculation of Hole Flow for Arbitrary Hole Concentration

In the text it was assumed that the concentration of holes was sufficiently small so that the first term in the brackets of Eq. (26) could be neglected in comparison with unity, yielding Eqs. (28) and (29). We give

here the general integration of Eqs. (26) and (27) for p arbitrarily large. Equation (26) may be written in the form:

$$i_p = -\text{grad}\ \psi, \tag{1B}$$

where

$$\psi = kT\mu_p \left[\frac{2bp}{b+1} - \frac{b(b-1)N_f}{(b+1)^2} \log\left(1 + \frac{(b+1)p}{bN_f}\right)\right]. \tag{2B}$$

Equation (27) then becomes:

$$\nabla^2\psi = 0. \tag{3B}$$

The radial solution of this equation corresponding to a total current I_p is:

$$\psi = \psi_\infty + I_p/2\pi r. \tag{4B}$$

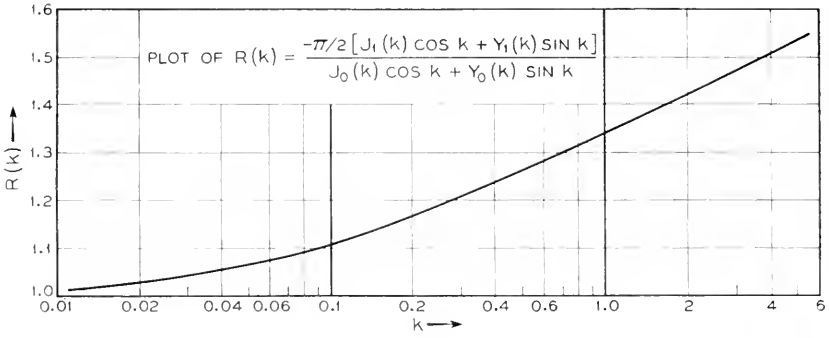

Fig. 10.—Correction factor for surface recombination.

The constants I_p and ψ_∞ are determined from the boundary conditions (30) and (31) of the text corresponding to $r = r_b$ and $r = \infty$. These conditions give:

$$\psi_\infty = kT\mu \left[\frac{2bp_i}{b+1} - \frac{2(b-1)N_f}{(b+1)^2} \log\left(1 + \frac{(b+1)p_i}{bN_f}\right)\right], \tag{5B}$$

$$I_p = 2\pi r_b\,(\psi(r_b) - \psi_\infty),$$

$$= 2\pi r_b \left[\frac{2b(p_b - p_i)}{b+1} - \frac{b(b-1)N_f}{(b+1)^2} \log \frac{bN_f + (b+1)p_b}{bN_f + (b+1)p_i}\right]. \tag{6B}$$

This equation is the appropriate generalization of Eq. (33) of the text. Since the equations are no longer linear, they do not apply strictly to the added hole concentration. However, if the normal hole concentration, p_0, is small, p_0 will be negligible in comparison with p_{ba} and p_a when the equa-

tions are not linear. Accordingly, to a close approximation, we may take for the added hole current:

$$I_{pa} = 2\pi r_b \left[\frac{2b(p_{ba} - p_a)}{b+1} - \frac{b(b-1)}{(b+1)^2} \log \frac{bN_f + (b+1)p_{ba}}{bN_f + (b+1)p_a} \right], \quad (7B)$$

which is the generalization of Eq. (34) of the text.

The value of p_{ba} and thus of I_{pa} may then be found by equating this expression with that of Eq. (3) for I_{pa}. This procedure yields the transcendental equation:

$$p_{ba} = -a \left[\frac{2b(p_{ba} - p_a)}{b+1} - \frac{b(b-1)N_f}{(b+1)^2} \log \frac{bN_f + (b+1)p_{ba}}{bN_f + (b+1)p_a} \right], \quad (8B)$$

where a is again defined by Eq. (6) of the text. This equation must be solved in general by numerical methods for a particular case. The equation simplifies for p_a either large or small compared with N_f. The latter case is treated in the text. The opposite limiting case of large hole concentrations is treated below.

For p_a large compared with N_f, the logarithm may be neglected, so that

$$p_{ba} = -2ab(p_{ba} - p_a)/(b+1). \quad (9B)$$

If, as in the text, it is assumed that a is small in comparison with unity, there results:

$$p_{ba} = 2abp_a/(b+1), \quad (10B)$$

and, using (3):

$$I_{pa} = -[2b/(b+1)]p_a kT\mu_p A/r_b. \quad (11B)$$

This differs from (7) by a factor $2b/(b+1)$. The equation corresponding to (8) will have this additional factor, and also the expression for the conductance, G, which, for large hole concentrations is:

$$G = G_0 + [2b/(b+1)](\alpha\beta\sigma_0 A/br_b)(p_a/n_0), \quad (12B)$$

in place of (38) of the text. Equation (16) which relates floating potential and conductance is general, and applies for arbitrary hole concentration.

Design Factors of the Bell Telephone Laboratories 1553 Triode

By J. A. MORTON and R. M. RYDER

(*Manuscript Received Aug. 3, 1950*)

IN DEVELOPING microwave relay systems for frequencies around 4000 megacycles, one of the major problems is to provide an amplifier tube which will meet the requirements on gain, power output, and distortion over very wide bands. As the number of repeaters is increased to extend the relay to greater distances, the requirements on individual amplifiers for the system become increasingly severe. A tube developed for this service is the microwave triode B.T.L. 1553, the physical and electrical characteristics of which were briefly described in a previous article.[1] In the development of such a tube, both theoretical and experimental factors are involved; illustration of these factors in some detail is the purpose of the present paper.

Given the application, a number of questions arise at the outset. What determines the tube type—why pick a triode for development, rather than a velocity variation tube, or perhaps a tetrode? What electrode spacings are necessary in such a tube, and what current must it draw? How is its performance rated, and how does it compare with other tubes? To what extent can the performance be estimated in advance? What experimental tests can give more precise information? Some answers to these questions were obtained by the use of figures of merit, which led up to the choice of a triode as most promising for development, and which also led to the subsequent method of optimizing the design for the particular system application of microwave amplifiers and modulators.

The design process may be said to proceed by the following series of steps:

1. Formulate the system requirements, frequently with the aid of one or more figures of merit. The purpose here is to concentrate attention upon the limitations inherent in the tube alone by eliminating considerations of circuitry or of other parts of the system. The figure of merit measures tube performance in an arbitrary environment, so chosen as to be simple, and also directly comparable to the actual system requirement.

2. Make tentative choices of tube type, and analyze further to find out

[1] J. A. Morton, "A Microwave Triode for Radio Relay," *Bell Laboratories Record* 27, 166–170 (May 1949).

how the figure of merit depends on the internal parameters of the tube, such as spacings, current density, and so on.

3. Optimize the internal parameters to make the figure of merit as good as possible, with due regard to practical limitations like cathode activity, life, cost, etc.

4. Use enough experimental checks to make sure the estimates are sound. Build then the type of tube which appears to fill the requirements best, including the practical as well as the technical limitations. The figures of merit serve now as quantitative checks both of how well the tube satisfies the application, and also of how accurate is the theory.

Given a good accurate design theory, the whole process could in principle be calculated in advance. Such a theory would permit great savings in effort, since spot checks of relatively few parameters are sufficient to insure accuracy even when the theory is used to predict a wide range of phenomena. The extent to which presently available microwave tube theory meets this need is considerable, as will appear from some of the results below.

The degree of accuracy required of a theory increases as the development process continues. For preliminary estimates, such as deciding what tube type to develop, the theory can be rather rough and still be satisfactory. For complete predictions of final performance, only experimental construction can suffice. By this means the theory can be checked, so that it can serve future designs with improved accuracy.

The method outlined here is not new, but rather follows standard practice fairly closely. It does, however, give more than usual quantitative emphasis to the figures of merit, using them to codify the procedure; and it incorporates a certain amount of quantitative calculation at microwave frequencies. It will be seen that the theory of Llewellyn and Peterson needs only some semi-empirical supplementation in the low-voltage input space, as has already been pointed out by Peterson.[2]

Preliminary Estimates—Choice of Tube Type

For the New York to Boston microwave relay, an output amplifier was developed using already available velocity-modulation tubes.[3] With four stagger-tuned stages, the amplifier proved satisfactory for this service, and in fact tests indicated that this system could be extended to considerably greater distances and still give good performance. It was apparent, however, that these amplifiers would not be satisfactory for a coast-to-coast system.

[2] L. C. Peterson, "Signal and Noise in Microwave Tetrodes," *I. R. E. Proc.* (Nov. 1947).

[3] H. T. Friis, "Microwave Repeater Research," *B. S. T. J.* 27, 183–246 (April 1948).

When this limitation became clear several years ago, a study was undertaken to determine which particular type of electron tube amplifier then known had the best possibilities of being pushed to greater gain-band products. The results of this study indicated that a very promising prospect was to build, for operation at 4000 megacycles, an improved planar triode, that is, one in which the active elements are on parallel planes.

In arriving at this conclusion, two general types of device were considered: velocity-modulated, as in a klystron, and current-modulated, as in a triode. (Nowadays, such a study would of course include traveling-wave tubes.) The conclusions were reached with the aid of the gain-band figures of merit, along the following lines:

Gain-Band Product

The system performance requirements demand amplifiers capable of reasonable gains and power outputs over prescribed bandwidths. However, it is known that bandwidth can be increased by complicating the circuits (double-tuning, stagger-tuning, etc.). Such factors, being common to whatever tube may be used, are extraneous to a discussion of tube performance, and accordingly the tubes are rated by their performance with simple, synchronous resonant circuits. Furthermore, even then the bandwidth can be increased at the expense of a corresponding reduction of gain, by simply depressing the impedance levels of the interstages. Since the product of gain and bandwidth remains constant, it is a suitable figure of merit, independent of the particular choice of bandwidth, provided the definition of gain is suited to the device.

Unfortunately there is more than one possible gain-band product, the appropriate form depending on how many simple resonant circuits shape the band of the amplifier stage. For example, a conventional pentode or a velocity-variation tube is usually used in conjunction with two high-Q resonant circuits, one each on input and output. If these are adjusted to give the same Q, then it is well known that, no matter what the bandwidth, the product of voltage gain and bandwidth is constant. (See Appendix 1)

$$| \Gamma_0 | B = | Y_{21} | / 2\pi \sqrt{C_{in} C_{out}} \tag{1}$$

Here Γ_0 is the mid-band voltage gain, B the bandwidth 6 db down (3 for each circuit), Y_{21} the stage transadmittance, and C_{in} and C_{out} the total effective capacitances of the resonant circuits, including the contributions of the tube*. It is assumed that the stage is matched into transmission lines of some suitable constant admittance level G_0.

In amplifiers using triodes such as the B.T.L. 1553 (or tetrodes) in

* As shown in Appendix 1, all quantities in equations (1) and (2) are the values effective at the electrodes adjacent to the electron stream.

grounded-grid circuits, the situation is different because the Q of the input circuit is always very much smaller than that of the output. Here a figure of merit independent of bandwidth is obtained from the product of power gain and bandwidth:

$$|\Gamma_0|^2 B = |Y_{21}|^2/4\pi G_{in} C_{out} \qquad (2)$$

Here G_{in} is the total conductance of the input circuit, including tube contributions*. The gain is again measured with the tube matched at an arbitrary admittance level G_0. The band, being now limited by only one tuned circuit, is somewhat different in shape from the above, and is taken 3 db down.

While each figure of merit gives an unequivocal rating of tubes of appropriate type, the intercomparison of the two types still depends on the bandwidth. In particular, as the band is widened, the two-circuit type (klystron) loses gain at the rate of 6 db per octave of bandwidth, while the one-circuit type (triode) loses only 3 db per octave. Consequently, if the two devices start with equal gains at some narrow bandwidth, the triode rapidly pulls ahead in gain as the bandwidth is increased.

The figure of merit equation (1) states that improved klystron performance implies either an increase in transadmittance Y_{21} or a decrease in the band-limiting capacitances C_{in} or C_{out}. According to the simplest klystron bunching concept,[4] the transconductance of such a tube may be increased indefinitely simply by making the drift time longer. Unfortunately, this simple kinetic picture does not take account of the mutually repulsive space-charge effects which set an upper limit to the useful drift time by debunching the electrons.[5] For a 2000-volt beam in the 4000-megacycle range, this limit is approximately three micromhos per milliampere. The 402A tube used in the New York to Boston system has already approached this limit within a factor of two. Since the capacitances are also quite small, the prospect is quite dubious for any considerable improvement in gain-band merit if the simple klystron type of operation were to be used.

Improvements are possible in a klystron by changing the manner of operation so as to lower the drift voltage V_0, because the aforesaid transadmittance limit is proportional to V_0^{-n}.* This prospect is also relatively unattractive. To get transadmittance values anywhere near the triode would require low voltages and close spacings somewhat like the latter, and would encounter space-charge difficulties involved in handling a large current in a low-voltage drift space. Furthermore, the tube would be more

[4] D. L. Webster, *Jour. App. Phys.* 10, 501–508 (July 1939).
[5] S. Ramo, *Proc. I. R. E.* 27, 757–763 (December 1939).
* The value of n may vary between $\frac{1}{4}$ and $\frac{3}{4}$. See reference 5.

complex, having several grids instead of one. A number of modifications of klystron operation were considered, but all looked more complex mechanically and more speculative theoretically than a triode.

In a triode there is also an upper limit to the transconductance that can be achieved by spacing cathode and grid more closely. This limit would be reached if the spacing were so close that the velocity produced by the grid voltage were of the same order as the average thermal velocity of cathode emission. The triode limit of some 11,000 micromhos per milliampere is, however, many times greater than that for ordinary klystrons. What is still more important is the fact that previous microwave triodes were still a factor of twenty to twenty-five below this limit, leaving considerable room for improvement. Thus, if mechanical methods could be devised for decreasing the cathode-grid spacing and at the same time maintaining parallelism between cathode and grid, it seemed highly probable that great improvements would be available from a new triode.

The choice to develop a triode for this application was therefore taken not merely on the basis of simplicity, but also with the expectation that performance improvements would be not only larger but also more certainly obtainable than by use of a modified klystron. Moreover, the possibilities of using the triode over a wide frequency range in other ways—as a low noise amplifier, modulator and oscillator—lent additional weight to its choice. By translating the known requirements on gain, bandwidth and power output into triode dimensions as discussed below, it was found that the input spacings of existing commercial tubes would have to be reduced by a factor of about five. In addition, cathode emission current densities would have to be increased about three times. A design was evolved in which the required close spacings could be produced to close tolerances by methods consistent with quantity production requirements. The B.T.L. 1553 tube was the result (Fig. 1). Many of its design features were adopted for use in the Western Electric 416A tube, which is an outgrowth of this investigation.

Description of B.T.L. 1553 Triode*

The electrode spacings of this tube and of a 2C40 microwave triode are shown in Fig. 2. In the 1553, the cathode-oxide coating is .0005″ thick, the cathode grid spacing is .0006″, the grid wires are .0003″ in diameter, wound at 1000 turns per inch, and the plate-grid spacing is .012″. It is interesting to note that the whole input region of the 1553 including the grid is well within the coating thickness of the older triode.

The arrangement of the major active elements of the tube is shown in

* This section is repeated from reference 1 for completeness.

Fig. 3. This perspective sketch has been made much out of scale so that the very close spacings and small parts would be seen. The nickel core of the cathode is mounted in a ring of low-loss ceramic in such a manner that the nickel and ceramic surfaces may be precision ground flat and coplanar. A thin, smooth oxide coating is applied to the upper surface of

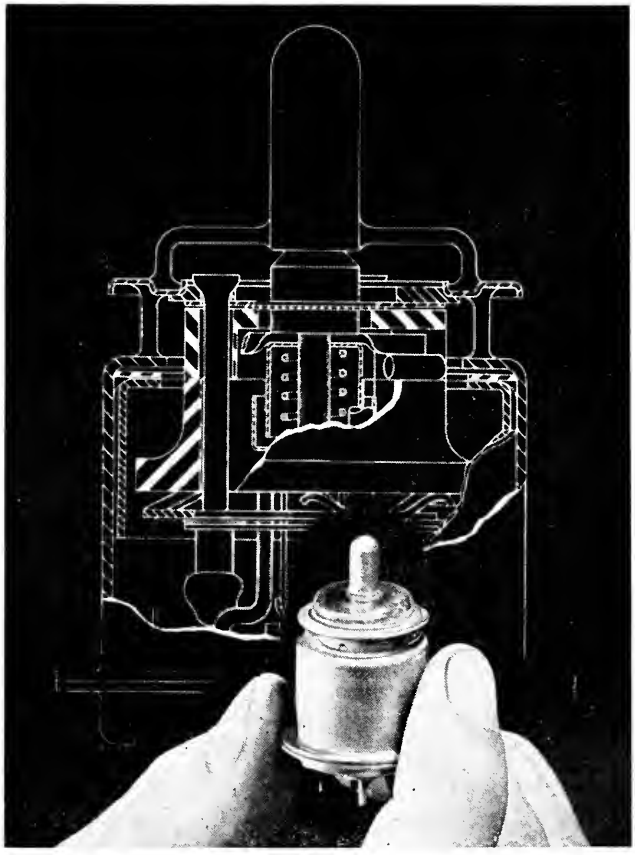

Fig. 1.—The B.T.L. 1553 microwave triode with a cross-section drawing of it in the background.

the cathode by an automatic spray machine developed especially for this tube. With this machine, a coating of 0.0005″ ± 0.00002″ may be put on under controlled and specifiable conditions. To insure long life with such a thin coating, it was necessary to develop coatings from two to four times as dense as those used in existing commercial practice.

The grid wires are wound around a flat, polished molybdenum frame

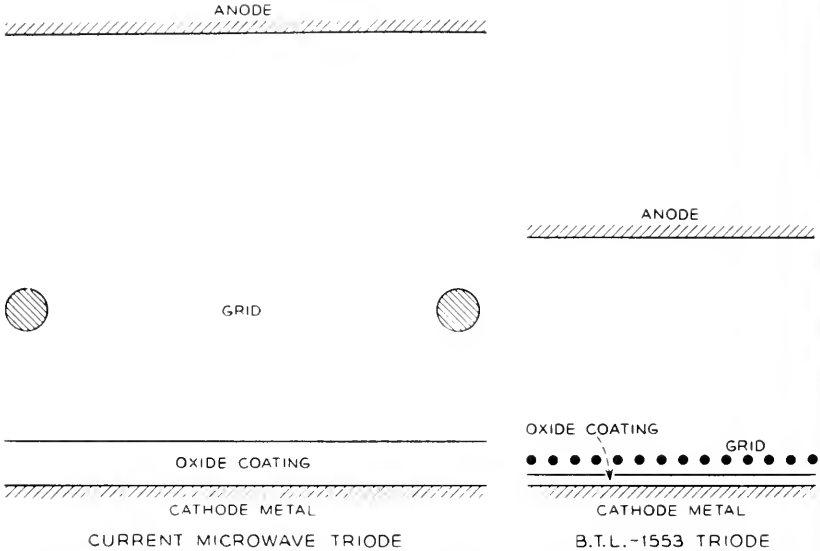

Fig. 2.—Comparison of the spacings of the 1553 triode at the right with a previously existing microwave triode at the left.

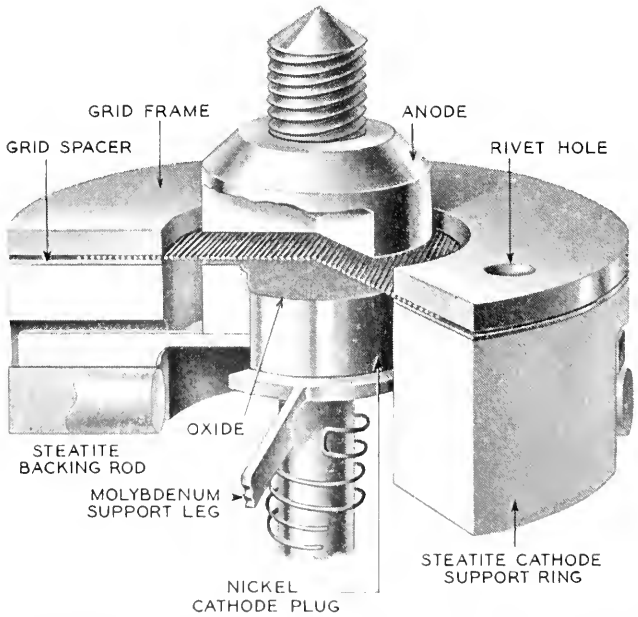

Fig. 3. Perspective drawing of the active elements of the 1553 close-spaced triode.

that has been previously gold sputtered. The winding tension is held within ± 1 gram weight to about 15 gram weight, which is about sixty per cent of the breaking strength of the wire. This is accomplished by means of a small drag-cup motor brake, a new method which was developed especially for these fine grids. The grid is then heated in hydrogen to about 1100°C, at which point the gold melts and brazes the wires to the frame. The mean deviation in wire spacing is less than about ten per

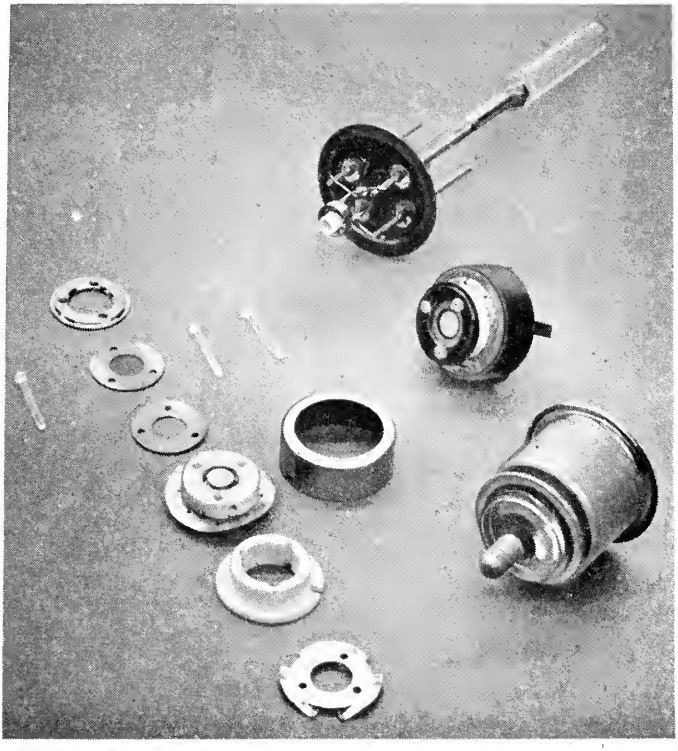

Fig. 4.—Physical appearance of the elements comprising the 1553 triode.

cent, and in fact these grids are fine enough and regular enough to be diffraction gratings as is shown in Fig. 5. In this figure, a fourth order spectrum diffracted by one of these grids can be seen. The third order, which should be absent because the wire size is about one-third of the pitch, is much less intense than the fourth. Proper spacing of the grid is then obtained by a thin copper shim placed between the cathode ceramic and the grid frame. Its thickness must be equal to the coating thickness, plus the thermal motion of the cathode, plus the desired hot spacing.

The cathode, spacer, and grid comprising the cathode-grid subassembly are riveted together under several pounds of force maintained by the molybdenum spring on the bottom of the assembly. The rivets are three synthetic sapphire rods fired on the ends with matching glass. In Fig. 4, the parts comprising this assembly are shown in appropriate pile-up sequence at the left, and the completed cathode-grid subassembly is shown at the right between the bulb and the press. The grid-anode spacing of .012″ is easily obtained by means of an adjustable anode plug the surface of which is gauged relative to the bulb grid disc.

Fig. 5.—Spectrum formed by the grid of the 1553 microwave triode.

TABLE I

LOW-FREQUENCY CHARACTERISTICS

For $V_p = 250$ V, $I_p = 25$ ma, $V_g = -0.3$ V

$g_m = 50{,}000$ μmhos	$C_{kg} = 10$ $\mu\mu f$
$\mu = 350$	$C_{gp} = 1.05$ $\mu\mu f$
$r_p = 7000$ ohms	$C_{kp} = .005$ $\mu\mu f$

The higher current density of 180 milliamperes per square centimeter, the thin dense cathode coating, and the very close spacings, posed a problem in obtaining adequate emission and freedom from particle shorts, and had to be solved by quality control methods because of the large number of factors involved and the precision required. Tubes, subassemblies, and testers have been made in batches and studied by statistical methods. To achieve a state of statistical control on emission, and freedom from dust particles, it is necessary to process the parts and assemble the tubes in a rigorously controlled environment. Completely air-conditioned processing and assembly rooms operating under rigorous controls have been found necessary[6]. Under such controlled conditions, good production yields with satisfactory cathode activity have been obtained,

[6] R. L. Vance, *Bell Laboratories Record*, 27, 205–209 (June 1949).

whereas without such conditions not only was the yield low but it was difficult to ascertain just what factors were operating to inhibit emission and to cause cathode-grid shorts.

A summary of the pertinent low-frequency characteristics of the 1553 triode is given in Table I. It should be noticed that, at plate currents of 25 milliamperes, the transconductance per milliampere is about 2000, that is, about one-fifth of the theoretical upper limit. At lower currents this ratio is higher: at 10 milliamperes, for example, it is 3000 micromhos per milliampere. Diodes with the same spacings have about twice these values of transconductance per milliampere, showing that the grid is fine enough to obtain fifty per cent of the performance of an ideal grid.

Triode Design Requirements

Analysis of the figure of merit can well begin by devoting attention to the band-limiting capacitance C_{out} of the output circuit. First, some question may be raised as to the applicability of the concept of a simple *L-C* shunt resonant circuit at high frequencies, where the circuit parameters are actually distributed, not lumped. Suppose the actual circuit admittance is $Y_x = G_x + jB_x$. In order to represent it as a simple shunt resonant circuit of admittance $Y_p = G_p + j\omega C_p + 1/j\omega L_p$, we need only require that the two be equal and have equal derivatives with respect to frequency at the center frequency $f_0 = \omega_0/2\pi$. Accordingly the "effective values" of the actual admittance are given by the following equations:

$$\begin{aligned} G_p &= G_x\,(\omega_0) \\ C_p &= \tfrac{1}{2}(B_x' + B_x/\omega_0) \\ \frac{1}{L_p} &= \tfrac{1}{2}(\omega_0^2\, B_x' - \omega_0 B_x) \end{aligned} \tag{3}$$

From this development one sees that the representation neglects G_x', the first derivative of the conductance, but otherwise is correct to first order as a function of frequency.

There are important cases where this representation as a simple circuit does not hold. For example, double-tuned circuits having two local resonances have a fundamentally different band shape. However, such complication of the circuits has been excluded from the figure of merit on the ground that it is purely a circuit "broad-banding" problem: having determined the performance of the tube for simple circuits, any broad-banding (double-tuning, staggering, etc.) will give a calculable improvement which does not depend upon the tube. Accordingly, to compare tubes it is sufficient to consider standard simple circuit terminations, tuned to the same frequency.

The total capacitance C_{out} includes two contributions: from the active electrode area inside the tube (C_{22}) and from the passive resonating circuit (C_{p2}). It is convenient to consider these separately, writing the figure of merit as follows:

$$|\Gamma_0|^2 B = \frac{|\Gamma_{21}^2|}{4\pi G_{11} C_{22}} \frac{1}{\left(1 + \frac{G_{p1}}{G_{11}}\right)\left(1 + \frac{C_{p2}}{C_{22}}\right)} \tag{4}$$

The first factor is the "intrinsic" electronic figure of merit of the active transducer alone, while the second factor expresses the deterioration caused by input passive circuit loss G_{p1} and output passive circuit capacitance C_{p2}, both of which should ideally be held as small as possible.

Consider the first factor, the intrinsic electronic gainband product which depends only upon the properties of the electron stream and the electrode dimensions in the regions occupied by the electron stream.

It is the responsibility of the tube design engineer to maximize this product consistent with any limitations which may be imposed by mechanical, emission, thermal or circuital considerations.

On the other hand, in maximizing this intrinsic gain-band product, the tube engineer must not proceed in ignorance of the effect of his actions on the possibility of obtaining a favorable value for the second factor. For example, he may attempt to make C_{22} so small (in order to maximize the first factor) that it becomes physically impossible to obtain an effective circuit capacitance C_{p2} which is not large compared to C_{22}. In such a case, the actual gain-band product would be much smaller than the intrinsic product of which the tube would be capable if circuit capacitance were negligible. Such a balancing of effects will become apparent from the subsequent discussion.

It is desired, therefore, to express the transadmittance, input conductance and output capacitance of the electronic transducer in terms of such parameters as cathode current density, electrode dimensions, frequency and potentials in such a way that it will become clear how a maximizing process may be carried out by adjusting these parameters.

As a first approximation let us use the results of Llewellyn and Peterson's analysis of plane-parallel flow[7], which makes the following assumptions:

1. All electrons are emitted with zero velocity.
2. All electrons in a given plane have the same velocity.

[7] F. B. Llewellyn and L. C. Peterson, "Vacuum Tube Networks," *Proc. I. R. E.*, 32, 144–166 (1944).

3. The dimensions of the grid are infinitesimal compared to the electrode spacings.
4. The electrode dimensions are small compared to the wavelength.

It can be shown that the intrinsic gain-band product may be expressed in the following two ways:

$$\begin{aligned} M_i &= K\left[\frac{1}{x_1}\right]\left[\frac{F_1^2(\theta_1)}{\theta_1 F_3(\theta_1)}\right]\left[\theta_2 F_2^2(\theta_2)\sqrt{V_p}\right] \\ &= K'\left[\frac{1}{j}\right]\left[\frac{F_1^2(\theta_1)}{\theta_1^4 F_3(\theta_1)}\right]\left[\theta_2 F_2^2(\theta)\sqrt{V_p}\right] \end{aligned} \tag{5}$$

where K, K' are parameters which are functions only of frequency.

x_1 is the cathode-grid spacing in cm

θ_1 is cathode-grid transit angle and $\theta_1 = \dfrac{126}{\lambda}\left(\dfrac{x_1}{j}\right)^{1/3}$

j = cathode current density in amp/cm^2

θ_2 = grid-anode transit angle and $\theta_2 = \dfrac{6300\ x_2}{\lambda\sqrt{V_p}}$

and $F_1(\theta_1)$, $F_2(\theta_2)$ and $F_3(\theta_1)$ are complicated functions of their respective transit angles.

Consider frequency to be given as part of the specifications on the tube.

Variation with Current Density, j

In the first formulation the current density is involved only in the second factor. This factor is a function only of $\theta_1 = \left(\dfrac{x_1}{j}\right)^{1/3}$ and is shown plotted in Fig. 6. If x_1 and λ are considered to be held fixed for the moment the first maximum at $\theta_1 \to 0$ requires j to be as large as possible consistent with emission limitations and life. For the 1553 the cathode current density is set at 180 ma/cm^2.

The other maxima at larger values of θ_1 (and smaller values of j), where $F_3(\theta_1)$ goes through zero, correspond to transit angles where $G_{11} \to 0$ in the single-valued velocity theory. These maxima cannot be taken at face value, however, to indicate maxima in the unequal-Q gain-band product since they violate the assumption that $Q_1 << Q_2$ for which the formula was developed. To make a study of gain-band variation in this region therefore entails a study of gain-band product as a function of bandwidth, as was pointed out previously in connection with comparison of the equal-Q and unequal-Q cases. Such maxima are of interest pri-

marily in narrow band cases so that for the present we shall concern ourselves only with the first maximum at $\theta_1 \rightarrow 0$ and j indefinitely large.

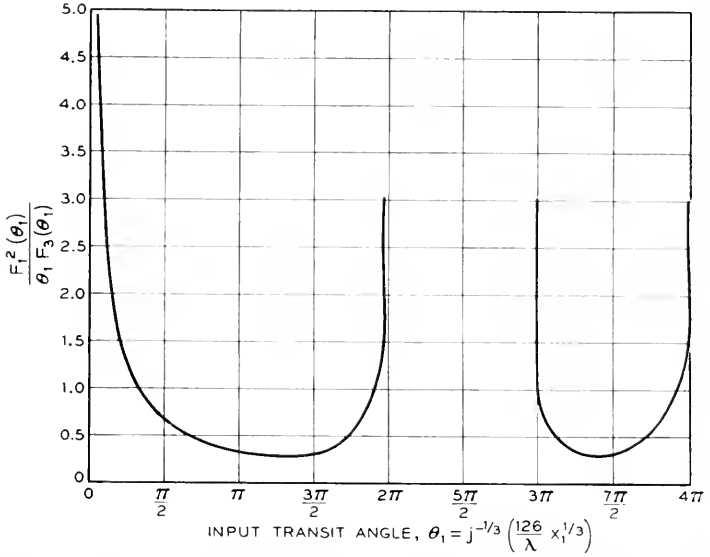

Fig. 6.—Gain-band product dependence on current density (j), with input spacing (x_1) fixed.

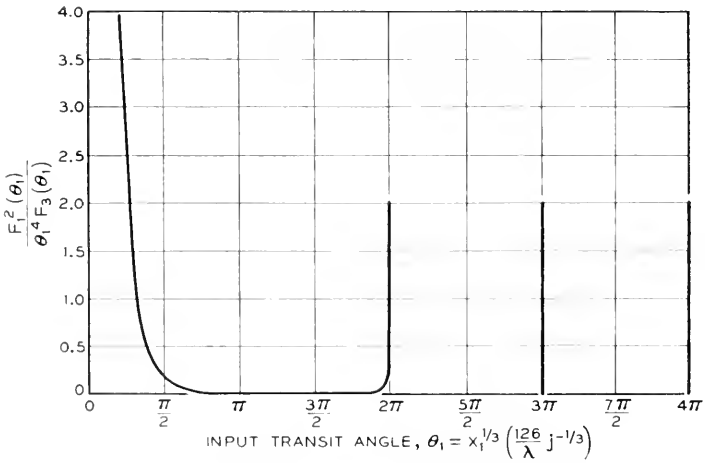

Fig. 7.—Gain-band product dependence on input spacing (x_1), with current density (j) fixed.

Variation with Cathode-Grid Spacing, x_1

Now consider that j has been fixed at the largest permissible value according to the previous section and consider the second formulation

for M_i. The spacing x_1 is involved only in the second factor which again is a function only of $\theta_1 = \left(\frac{x_1}{\lambda}\right)^{1/3}$. We again have a strong first maximum at $\theta \rightarrow 0$ requiring x_1 to be as small as possible (Fig. 7). Other maxima are indicated at larger values of θ_1 (and larger values of x_1) again at points where $G_{11} \rightarrow 0$ and the same remarks apply here as were made in the previous section. For broad-band optima we are therefore interested in minimum values of x_1.

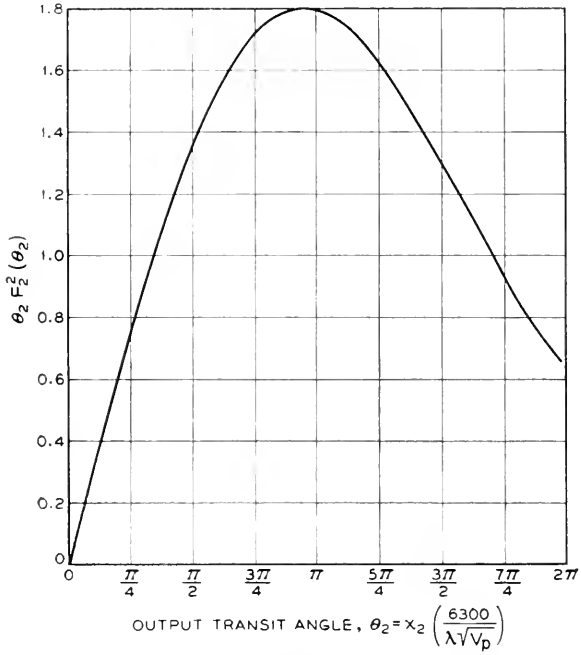

Fig. 8.—Gain-band product dependence on output spacing (x_2).

Variation with Anode-Grid Spacing, x_2

The anode-grid spacing x_2 is involved only in the third factor of either formulation. This factor is a function of output transit angle θ_2 and exhibits a maximum for $\theta = 2.9$ radians as shown in Fig. 8. This optimum at a fairly large value of θ_2 is due to the fact that the capacitance C_{22} varies as $1/x_2$ whereas the coupling coefficient of the stream to the gap decreases more slowly at first than the capacitance so that the ratio Y^2_{21}/C_{22} improves as the spacing becomes moderately wide. The optimum θ_2 corresponds to an optimum value of x_2 which of course depends upon the plate voltage and frequency of operation. For the 1553 at 250 volts and 4000 Mc/s, the optimum output spacing is .022″.

Limitations in Choosing Optimum Parameters

Generally, there are mechanical, thermal, emission and specification limits which prevent the realization of optimum values for all of the above parameters simultaneously. A good design is one in which a nice balance is effected between these various optima and their limitations.

Limitations on Emission Current Density, j

It is generally true that the life of a thermionic electron tube varies inversely as the average cathode current density in a complicated fashion. The maximum permissible value of j is therefore always a compromise between our desire for highest figure of merit and long life. In the present state of the cathode art as it has been evolved for the 1553 triode it is possible to operate at a current density of 180 ma/cm^2 and obtain an average life of several thousands of hours. It is perhaps of interest to note that it was necessary to develop much more dense and smooth oxide coatings in order to make possible such life in the thin coatings necessary for operation at such close spacings.

Limitations on Cathode-Grid Spacing, x_1

Consider the limitations in reaching the optimum in x_1. There is, of course, the obvious one that it is mechanically and electrically not possible at present to make x_1 equal to zero and still retain the essential features of unilateral controlled space charge flow. Granting then that the spacing cannot be zero, we must choose the smallest value of x_1 for which parallelism and reasonable tolerances can be maintained. To this end in the 1553 a value of $x_1 = .0006''$ is very near this limit with present structures.

There is, however, at present another limitation which is essentially mechanical in nature but makes itself felt electrically in a way not indicated in the above simplified theory. This theory has assumed that the grid dimensions are infinitesimally thin compared to the electrode spacings. However, if this is not the case then the grid has less control action than an ideal fine grid, and the intrinsic gain band product must be reduced by still another factor F_1 which is a function of the grid transmission factor $a = \frac{p - d}{p}$ and the ratio $\frac{x_1}{p}$ where p is the pitch distance between grid wire centers and d is the diameter of grid wires. This function has the form shown in Fig. 9.*

Thus if the grid pitch and wire diameter are mechanically limited to some finite though small values, the optimum in input spacing x_1 will

* Data transmitted informally from C. T. Goddard and G. T. Ford.

still be for $x_1 \rightarrow 0$ but will not increase so strongly as $x^{-4/3}$ as before but much more slowly, about as $x^{-1/3}$. The grid dimensions should consequently be made as small as possible while still maintaining a transmission fraction at no less than 0.5 and at the same time not allowing mean deviations in pitch more than about 15%.

In the 1553 our best grid techniques today have led to a stretched grid (which does not move appreciably during temperature cycling) having a transmission factor of approximately 0.7, a pitch distance of .001″ and a mean deviation in pitch of less than 15%. For such a grid further decreases in input spacing without refining the grid will not pay off very rapidly, since we are on the maximum slope portion of the function F_4.

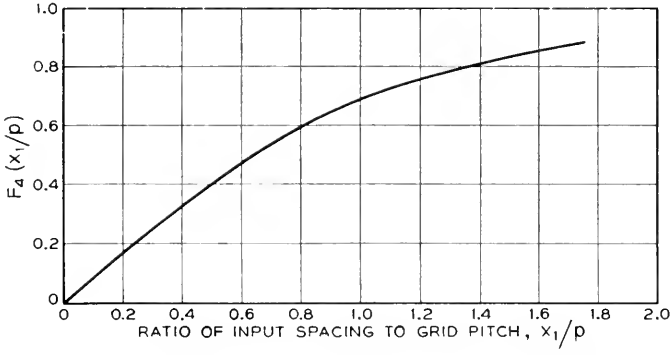

Fig. 9.—Dependence of gain-band product on grid pitch.

Limitations on Anode-Grid Spacing, x_2

In considering the choice of output spacing we must attain a balance among the following considerations:

a. The optimum transit angle $\theta_2 = 2.9$ radians requires a spacing which varies with plate voltage and with frequency. For 250 volts and 4000 Mc/s, this optimum is .022″.

b. The anode heat dissipation must be closely watched because the glass seal in this type of tube is very close to the anode. For the 1553, a maximum of 50 watts per square centimeter of anode active surface is safe. With a maximum cathode current density of 180 ma/cm², set by life considerations, heat dissipation limits the plate voltage to 275 volts unless the current is lowered.

c. If the anode is moved too far out, keeping its voltage constant, then in order to draw the desired current the grid must go positive, perhaps drawing excessive grid current. The grid shielding factor μ cannot be reduced without harming the transadmittance and feed-

back values; accordingly the cathode current would have to be reduced below the maximum permissible from life considerations.

d. The circuit degradation factor $(1 + C_{p2}/C_{22})^{-1}$ becomes more unfavorable as the active capacitance C_{22} is reduced by widening the output spacing. For discussion and calculation of this factor, see Appendix 2.

e. A wider output spacing, by virtue of the reduced capacitance, permits a higher maximum frequency limit on the tube.

The actual choice of output spacing in the 1553 is .012″. This compromise between the foregoing factors appears to be suitable at 4000 Mc/s. The output transit angle of 1.6 radians gives 78% of the theoretical optimum intrinsic gain-band product. The anode dissipation is near the maximum safe value for the maximum allowable cathode current. The grid runs very close to cathode potential so that grid current is small. The circuit degradation factor has a value of about 0.8, while the upper frequency limit of the tube is satisfactory (about 5000 Mc/s).

The optimum design just described is an attempt to get the best possible gain-band product in the resulting tube, and is based on a particular electronic theory (that of Llewellyn and Peterson). Two points remain to be discussed. (1) What would be the result of optimizing for other merit figures such as power-band product or noise figure, and (2) how valid is the theory?

Power-Band Product

The radio relay amplifier requires not only gain, but perhaps even more, power output. In such a case, the design specification of greatest importance is the bandwidth over which a certain power output can be obtained with a specified maximum distortion, and is expressed by an analogous figure of merit, the power-band product.

Of the many methods of specifying distortion, one which is particularly useful in this connection is the "compression", that is, the amount by which the gain is reduced from the small-signal value. In an amplitude-modulated system, the compression would be a direct measure of nonlinear amplitude distortion in the amplifiers. In the actual relay, using FM, compression is an indication that the amplifier is approaching its maximum limit of power output.

The maximum power output depends not only on how much current the tube can carry, but also on the magnitude of the load impedance into which this current works, which in turn depends upon the bandwidth of the load. To compare tubes without need of specifying any bandwidth, one notes that the product of power output and band-

width is a constant, a figure of merit. The derivation is outlined in Appendix 1.

$$P_0 \cdot B = \frac{I_{20}^2 F^2(C) F_2^2(\theta_2)}{4\pi C_{\text{out}}} \tag{6}$$

The numerator here is just the square of the maximum ac current; that is, the dc current I_{20}, multiplied by a factor $F(C)$ depending on the allowable compression C, and by the gap coupling coefficient $F_2(\theta_2)$ of the electron stream to the output gap. The latter is of course a function of the output transit angle θ_2. It is assumed that the load is a matched simple resonant circuit and the band is taken 3 db down.

The power optimum must clearly be somewhat different from the gain optimum previously discussed. For example, the transadmittance does not appear here, nor does any property of the input circuit; while the magnitude of the direct electron current, which did not appear in the gain-band product, is now important. The capacitance of the output circuit appears in both figures of merit.

In terms of internal parameters of the tube, application of Llewellyn and Peterson's theory along the lines previously discussed leads to the following expression for power-band product:

$$M_i\,(P) = K[Aj^2\, F^2(C)]\; [\theta_2\, F_2^2\,(\theta_2)\, \sqrt{V_p}] \tag{7}$$

where A is the electrode area, $F^2(C)$ is a function of the allowable distortion limits, K is a constant which may depend upon frequency, and the other symbols are as before.

Considering first the dependence on output transit angle and plate voltage, one sees that this figure of merit has exactly the same form as the gain-band product. It is, however, not quite safe to assume therefore that exactly the same output configuration is still optimum, because the factors entering into the choice of output spacing have not exactly the same relative importance any longer; for example, a positive grid may be less objectionable, or a higher plate voltage may be permissible. Still, as a first approximation one may assume the output configuration to be already somewhere near optimum.

Other factors of the power-band figure of merit show considerable difference from the gain-band product. For instance, the electrode area enters the picture explicitly, suggesting that a larger area tube would give more power. The current density enters squared instead of only to the $\frac{2}{3}$ power; the explicit dependence on input spacing is missing. The compression function $F(C)$ depends mostly on the input conditions in a complicated way difficult to calculate. It can be approximated graphically from static characteristics.

A power tube similar to the 1553 might therefore be larger in electrode area, might have a coarser grid and wider input spacing, and perhaps would differ somewhat in output configuration, particularly if the plate voltage were raised. Any cathode development permitting a higher current density would improve the power output more than the gain, and might well lead to a drastic anode redesign to permit larger plate dissipation.

Similarly, a design to optimize noise figure would lead to still a third version of the tube, in which one might consider such things as critical relationships between input and output spacings.

For the 1553 at 4000 megacycles the following quantitative data may be quoted in order to check the gain-band product estimates.[8]

$$| Y_{21} | = 39.10^{-3} \text{ mhos}$$

$$G_{11} = 73 \cdot 10^{-3} \text{ mhos}$$

Note that the transadmittance is less than the dc value of $45 \cdot 10^{-3}$ mhos by only about 15%, while the input conductance, instead of being equal to the transadmittance as at low frequencies, is almost twice as large, on account of loading of the input gap by electrons returning to the cathode. Using the active capacitance C_{22} of .477 $\mu\mu f$, the intrinsic gain band product is:

$$\Gamma \cdot B = Y_{21}{}^2/4\pi G_{11} C_{22} = 3480 \text{ megacycles.}$$

With the somewhat optimistic capacitance degradation factor of .81 computed in Appendix 2, the gain band product would be reduced to 2820 megacycles.

The experimental average value is about 1100 megacycles.[9] The difference is probably due in part to resistive loss in the passive input circuit, which may be calculated as follows: Neglecting feedback, the input circuit may be represented as containing a resistance R_s in series with the short-circuit input admittance $g_{11} + jb_{11}$. Robertson gives the following values for these elements:

$$g_{11} = 73 \cdot 10^{-3} \text{ mhos}$$

$$b_{11} = 26 \cdot 10^{-3} \text{ mhos}$$

$$R_s = 7.6 \text{ ohms}$$

Accordingly, the input degradation factor $R_{11}/(R_{11} + R_s)$ should be $11.2/(11.2 + 7.6) = .60$, giving a computed overall gain-band product of 1690 megacycles. The best tubes sometimes exceed this figure. Tubes

[8] S. D. Robertson's measurements at 4000 megacycles, *B. S. T. J.*, 28, 619–655 (October 1949).

[9] A. E. Bowen and W. W. Mumford "Microwave Triode as Modulator and Amplifier," this issue of *B. S. T. J.*

with lower values may have excessive input circuit loss or may have narrower bandwidth on the input side than has been assumed. Further measurements, by elucidating this point, might lead to a better design of tube and circuit.

An entirely similar calculation can be made for the power-band product. The additional assumptions required are that the compression function $F^2(C)$ has the conservative value of $\frac{1}{2}$, and the output coupling coefficient $F_2(\theta_2)$ is taken as 0.9. The power-band product at 4000 megacycles is then computed to be 50 watt megacycles, which is quite close to the figures found by Bowen and Mumford.

Refinements of the Electronic Theory

In the electronic computations above, the single-valued theory was used because it is the simplest theory which describes the high frequency case at all accurately. The most important discrepancy between the rigorous theory and the actual situation is the first theoretical assumption listed above, that the electrons are emitted from the cathode with zero velocity. For actual cathodes the velocity of emission is not zero nor uniform but has a Maxwellian distribution such that the average energy away from the cathode is $\frac{1}{2}\ k\ T_k$, or about equivalent to the velocity imparted by a potential drop of 0.04 volt for an oxide cathode at 1000°K. There result several effects whose general nature is known but which have not yet been formulated into a rigorous quantitative theory valid at high frequencies.

(1) A potential minimum is formed at a distance on the order of .001″ in front of the cathode instead of at the cathode as in the simple theory. This distance is not negligible for close-spaced tubes; so that, for very close spacings, even perfect "physicists' grids" approach a finite trans-conductance limit. [van der Ziel, Philips Research Reports 1, 97–118 (1946); Fig. 2.]

(2) Because the potential minimum implies a retarding field near the cathode many electrons emerging from the cathode are forced to return to it. These returning electrons absorb energy from the signal and also induce excess noise in it, both effects becoming important at high frequencies.

The effects of initial velocities on the figures of merit can be measured experimentally. For example, the circuit and electronic impedances of diodes and triodes at 4000 Mc have been measured by Robertson.[8] Such measurements can determine the electronic loading and noise separately from the circuit degradation effects and are therefore a highly effective

[8] loc. cit.

method of circuit design as well. Robertson found that the input circuit structure of the 1553 produces a measurable impairment in its gain-band product, which redesign of both tube and circuit may be able to improve. Comparison of his results with the theory has given a better understanding of the limits of high-frequency performance, and has lent some support to the following set of rules of thumb which have been in use for some time:

1. The input loading arising from the returning electrons is considerable, the input conductance of these tubes at 4000 Mc being about double the theoretical value of Llewellyn and Peterson.[7]
2. The input noise of these close-spaced tubes checks well with what one would expect of a low-frequency diode with Maxwellian velocities, whose solution is known. In high-frequency noise calculations, therefore, one can use with some confidence Rack's suggestion that cathode noise can be regarded as an effective velocity fluctuation at the virtual cathode.[10]
3. Single velocity theory seems to hold well when velocities are much larger than Maxwellian, drift times are not more than a few cycles, electron beams are short compared to their diameter, and no exact cancellations of large effects are predicted. In particular it holds well for the 1553 output space and for calculations of the high-frequency trans-admittance.

Extensive calculations of signal and noise behavior in planar multigrid tubes have been made by L. C. Peterson, using the single-velocity theory except for an empirical value of input loading, and using Rack's suggestion for cathode noise.[11] The results so far checked have agreed well with experiment.

In short, the optimum design for the tube is still given fairly closely by the figures of merit based on the approximate theory, but the performance will fall somewhat short of the predictions of the simple theory; performance can be estimated with the aid of the experimental measurements and rules of thumb just described.

Summary

From the foregoing calculations we draw a number of conclusions:

1. The figures of merit can be validly analyzed into their dependence on more elementary properties like transadmittance, circuit capacitance, input loss resistance, and so on.

[7] loc. cit.

[10] A. J. Rack "Effects of Space Charge and Transit Time on the Shot Noise in Diodes," *B. S. T. J.*, 17, 592–619 (October 1938).

[11] L. C. Peterson "Space Charge and Noise in Microwave Tetrodes," *Proc. I. R. E.*, 35, 1262–1274 (November 1947).

2. Even rough calculations, such as the coaxial line approximations used in Appendix 2 are close enough to the facts to indicate whether the design is close to an optimum with respect to such parameters as output spacing, anode diameter, grid diameter, and the like. More accurate calculations and experiments can give more precise answers to these questions.
3. Some considerations such as cathode activity, tube life, heater power and so on have not yet been included in the analysis. However, systematic optimization for such parameters as are treated quantitatively is greatly facilitated. In general, each different figure of merit leads to a somewhat different optimum and hence a different version of the tube.

The design of tubes by the method of figure of merit has been outlined. The method is very general, but in essence has just three steps:

1. Formulate the system performance of the projected device with the aid of a figure of merit.
2. Find how the figure of merit depends upon the parameters of the tube, such as spacings, current, etc.
3. Adjust the tube parameters, subject to physical limitations, to optimize the figure of merit.

Acknowledgments

The development of this microwave triode has required not only the expert and highly cooperative services of a large team of electrical, mechanical, and chemical engineers but also the indispensable assistance of skilled technicians, all of whom worked smoothly together to develop these new materials and techniques to a point where they are specifiable and amenable to quantity production. It is not practical to mention all those who have made significant contributions to this development. The contributions of A. J. Chick, R. L. Vance, H. E. Kern and L. J. Speck, however, are of such outstanding nature that mention of them cannot be omitted.

APPENDIX 1

Derivation of the Figures of Merit

Gain-Band Figure of Merit

Let the problem be stated as the design of an amplifier tube to operate with as large gain over as wide a frequency band as practicable. As a standard environment, we use a single-stage amplifier working between equal resistive impedances. For three reasons this standard is suitable: it is simple; it corresponds closely to practicality in many cases especially

in the microwave field; and in most cases, it turns out that performance is limited by the same transadmittance to capacitance ratios as apply when the source and load impedances are not purely resistive. The terminology of high frequencies will be used but the analysis applies at all frequencies under the conditions stated.

Consider the over-all single-stage amplifier of Fig. A1-1 consisting of input resonator, tube and output resonator, to be a single transducer

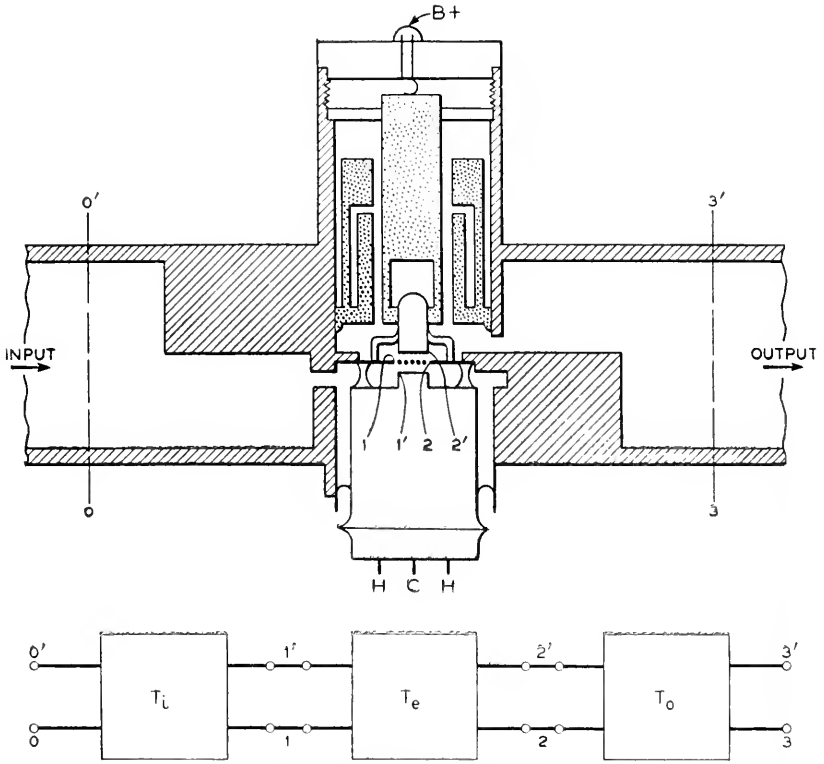

Fig. A1-1.—Microwave triode amplifier.

whose gain and bandwidth we wish to relate to the geometry and other pertinent characteristics of the circuits, bulb and electrode characteristics.

It is instructive to consider the whole transducer to be made up of three transducers in tandem as follows:

1. The input passive transducer, extending from the externally available input terminals (perhaps located somewhere in the driving wave guide or coaxial line) up to the internal input electrodes right at the boundary of the electron stream. Call this transducer T_i; in the

case of the grid-return triode of Fig. 1 it begins somewhere in the input wave guide at 0-0′ where only the dominant wave exists, includes the input external cavity and that portion of the tube interior right up to but not including the cathode-grid gap adjacent to the electron stream at 1-1′.

2. The output passive transducer, extending from the externally available output terminals located in the output wave guide through the output part of the bulb right up to the internal output electrodes at the boundary of the electron stream. Call this T_0; in the triode it

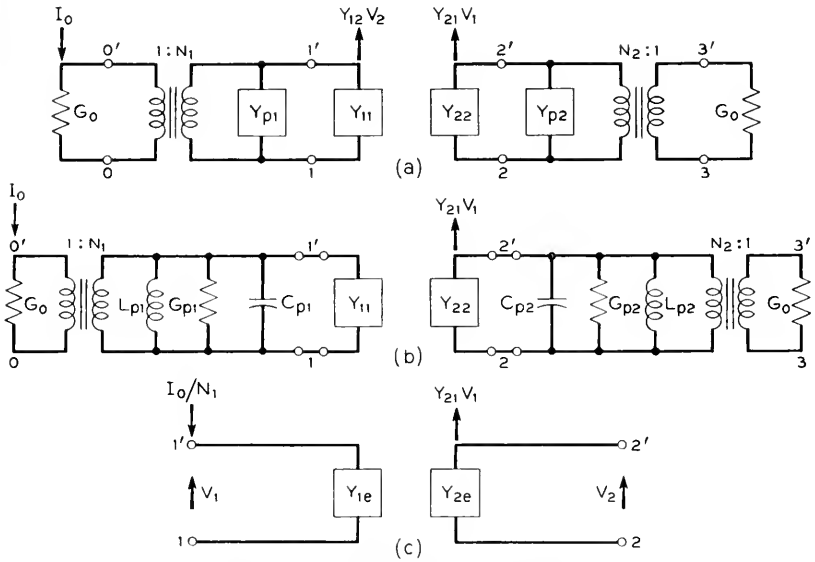

Fig. A1-2.—Amplifier representations.

extends from somewhere in the output wave guide at 3-3′ where only the dominant wave exists, includes the external coupling window, resonator cavity and output portion of the bulb, right up to the grid-anode gap adjacent to the electron stream at 2-2′.

3. The active electron transducer enclosing everything between the internal terminals of the above two passive coupling transducers—call this T_e—in the triode it extends from the cathode-grid gap adjacent to the electron stream at 1-1′ to the grid-anode gap adjacent to the electron stream at 2-2′. Geometrically it includes the stream and active portions of the electrodes. The term "active" will be applied to the electron stream and to those portions of the electrodes which interact directly with the stream.

We may represent these three transducers as in Fig. A1-2a, where the input and output transducers have each been replaced by an ideal transformer of turns ratio N and a shunt admittance Y_p. This representation is general enough for present purposes, provided that Y_p and N are allowed to be complex functions of frequency and provided that terminals 0-0′ and 3-3′ are chosen so that a potential minimum occurs at those points when points 1-1′ and 2-2′ are shorted.

The short-circuit admittances for the whole transducer as seen at terminals 0-0′ and 3-3′ are then

$$\begin{aligned} Y^*_{11} &= N_1^2 (Y_{11} + Y_{p1}) \\ Y^*_{22} &= N_2^2 (Y_{22} + Y_{p2}) \\ Y^*_{21} &= N_1 N_2 Y_{21} \\ Y^*_{12} &= N_1 N_2 Y_{12} \end{aligned} \tag{A1-1}$$

where the Y_{ij} are the short-circuit admittances of the electron transducer alone as seen at terminals 1-1′ and 2-2′.

If the feedback admittance Y_{12} is assumed negligible the insertion voltage gain may be written as

$$\Gamma(\omega) = \frac{2N_1 N_2 Y_{21}}{G_0(1 + \sigma_1)(1 + \sigma_2)}$$

where the sigmas are admittance-matching factors:

$$\sigma_1 = \frac{N_1^2(Y_{11} + Y_{p1})}{G_0} = \frac{Y^*_{11}}{G_0}; \qquad \sigma_2 = \frac{N_2^2(Y_{22} + Y_{p2})}{G_0} = \frac{Y_{22}^*}{G_0} \tag{A1-2}$$

The gain is maximum when σ_1, σ_2 are minimum, i.e., when tube and circuits are resonant and losses are minimum.

We may rewrite this in terms of the total Y^*_{ij} as follows:

$$\Gamma(\omega) = \frac{2Y^*_{21}}{G_0(1 + \sigma_1)(1 + \sigma_2)} \tag{A1-3}$$

Many practical cases are well approximated by the more special representation of Fig. A1-2b, where the turns ratios of the ideal transformers are real and independent of frequency, and the shunt admittance consists of ordinary lumped constant circuit elements. The feedback admittance Y_{12} is neglected.

This representation as simple, lumped-constant elements holds very well for any admittance, even a distributed, cavity-type microwave circuit, or an electronic admittance, provided that the combined circuit has no series and only one shunt resonance near the frequency band in

question. The "effective values" of the actual admittance are given by equations (3) of the text, as follows:

$$\begin{aligned} G_p &= G_x\,(\omega_0) \\ C_p &= \tfrac{1}{2}\,(B_x' + B_x/\omega_0) \\ \frac{1}{L_p} &= \tfrac{1}{2}\,(\omega_0^2\,B_x' - \omega_0\,B_x) \end{aligned} \tag{A1-4}$$

Let the complete admittances across nodal pairs 1-1′ and 2-2′ be called Y_{1e} and Y_{2e} as in Fig. A1-2c, which is an abbreviation of Fig. A1-2b from the point of view of the active transducer.

$$\begin{aligned} Y_{1e} &= G_1 + G_{p1} + G_{11} + j\omega C_{p1} + j\omega C_{11} + \frac{1}{j\omega L_{p1}} \\ &= G_{1e} + j\omega C_{1e} + \frac{1}{j\omega L_{1e}} \end{aligned} \tag{A1-5}$$

$$\begin{aligned} Y_{2e} &= G_2 + G_{p2} + G_{22} + j\omega C_{p2} + j\omega C_{22} + \frac{1}{j\omega L_{p2}} \\ &= G_{2e} + j\omega C_{2e} + \frac{1}{j\omega L_{2e}} \end{aligned}$$

where G_1 and G_2 are the line admittances as seen from the active transducer:

$$G_1 = G_0/N_1^2;\; G_2 = G_0/N_2^2.$$

The Q's of the circuit are defined as

$$\begin{aligned} Q_{1e} &= \omega_0\,C_{1e}/G_{1e} \\ Q_{2e} &= \omega_0\,C_{2e}/G_{2e} \end{aligned} \tag{A1-6}$$

The insertion voltage gain (2) may be written as follows to emphasize the manner in which it depends upon frequency:

$$\Gamma = \frac{2Y_{21}}{Y_{1e}Y_{2e}}\sqrt{\frac{G_{1e}G_{2e}}{(1+\mu_1)(1+\mu_2)}} \tag{A1-7}$$

Here $\mu = \sigma(\omega_0)$ is the matching factor at band center. Frequently the circuits are matched ($\mu_1 = \mu_2 = 1$) to avoid standing waves in system applications, and we shall discuss this case; but in any case μ_1 and μ_2 are constants with respect to frequency. For our standard circuits, G_{1e} and G_{2e} are independent of frequency; also ordinarily the transadmittance Y_{21} may be considered constant for bandwidths commonly encountered. There results then the fact that the voltage gain (and phase) depends on frequency in the same way as $(Y_{1e}\,Y_{2e})^{-1}$.

Since the gain varies with frequency, the amplifier will give approximately constant response only within a certain range of frequencies. The band of the amplifier is defined as that frequency interval within which the magnitude of the gain is constant within some specified tolerance; the bandwidth is the size of this interval. We wish to express the gain of the amplifier in terms of its bandwidth, in the following way:

The voltage gain of this amplifier has a maximum, called Γ_0, at band center frequency f_0. Take the band of the amplifier $B_N(A)$ as that interval within which the voltage gain is within a factor of 1/N times the maximum.

$$\left| \frac{\Gamma(\omega)}{\Gamma(\omega_0)} \right| \geq \frac{1}{N} \text{ defines } B_N(A) \tag{A1-9}$$

We can analogously define the band of a simple circuit $B_n(C)$ by the relation

$$\left| \frac{Y_{2e}(\omega_0)}{Y_{2e}(\omega)} \right| \geq \frac{1}{n} \text{ defines } B_n(C). \tag{A1-9}$$

It follows directly that

$$B_n(C) = \frac{G_{2e}}{2\pi C_{2e}} \sqrt{n^2 - 1}\,. \tag{A1-10}$$

Since the amplifier gain is inversely proportional to the product of the circuit admittances, it follows that $n_1 n_2 = N$.

The intrinsic bandwidth resulting from the tube admittance may not be suitable for the intended application. In that case the band may be widened by increasing G_{1e} or G_{2e} with a corresponding decrease in gain. We have then the problem of adjusting G_{1e} and G_{2e} for greatest band efficiency, i.e., maximum gain for a given bandwidth, with synchronous tuning. It turns out that if the bandwidth is less than that needed, then the circuit of higher Q should be lowered until either (a) the band becomes wide enough, or (b) the Q's become equal. In case (b), both Q's should then be lowered, maintaining equality, until the band is wide enough.

Two important limiting cases are to be considered: (a) $Q_{1e} = Q_{2e}$, i.e. the band is shaped equally by the input and output circuits; and (b) $Q_{1e} << Q_{2e}$, i.e. the band is shaped by only the output circuit. In the equal-Q case we have

$$\begin{aligned} \frac{G_{1e}}{C_{1e}} &= \frac{G_{2e}}{C_{2e}} \\ n^2 &= N \\ B_N(A) &= \frac{1}{2\pi} \sqrt{\frac{G_{1e}}{C_{1e}} \frac{G_{2e}}{C_{2e}}} \sqrt{N - 1}\,. \end{aligned} \tag{A1-11}$$

If only the output circuit is involved, then $N = n_2$ and the band of the amplifier, being shaped differently, is given by a different relation:

$$B_N(A) = \frac{G_{2e}}{2\pi C_{2e}} \sqrt{N^2 - 1}. \tag{A1-12}$$

In other words, a band shaped by only one circuit has the shape of (12), while a band shaped by two circuits has the shape (11). The maximum voltage gain is

$$| \Gamma_0 | = | \Gamma(\omega_0) | = \frac{2 | Y_{21} |}{\sqrt{G_{1e}G_{2e}(1 + \mu_1)(1 + \mu_2)}} \tag{A1-13}$$

Substituting for the G's in terms of the bandwidth, we have for the equal-Q case (from 11)

$$| \Gamma_0 | = \frac{| Y_{21} |}{2\pi\sqrt{C_{1e}C_{2e}}} \frac{2\sqrt{N - 1}}{\sqrt{(1 + \mu_1)(1 + \mu_2)}} \frac{1}{B_N} \tag{A1-14}$$

and for the unequal-Q case (from 12)

$$| \Gamma_0 | = \frac{| Y_{21} |}{\sqrt{G_{1e}}\sqrt{4\pi C_{2e}}} \frac{\sqrt{8}\sqrt[4]{N^2 - 1}}{\sqrt{(1 + \mu_1)(1 + \mu_2)}} \frac{1}{\sqrt{B_N}} \tag{A1-15}$$

These equations give the relationship between the gain and bandwidth of a transmission system shaped by two or one independent circuits, respectively. The comparison between these two cases is not quite straightforward. First, the band shapes (11) and (12) are different, although this difference is small enough to be ignored for $N < 2$ (6 db down). Second, the gain varies differently as the band is widened; the equal-Q case loses gain at 6 db per octave in bandwidth, the unequal-Q case only 3 db per octave. The comparison therefore depends on the bandwidth chosen. However, these formulas are still quite useful, especially in comparing two amplifiers of the same type or in optimizing an amplifier of one of the types.

From the equal-Q formula one notices that the product of insertion voltage vain and bandwidth does not depend on the bandwidth, but is a figure of merit by which two amplifiers of the same type (i.e. equal Q) but different gains and bandwidths can be compared. Since

$$C_{1e} = C_{11} + C_{p1}\ ;\ C_{2e} = C_{22} + C_{p2}$$

$$| \Gamma_0 | B_N = \left(\frac{Y_{21}}{2\pi\sqrt{C_{11}C_{22}}}\right)\left(\frac{1}{\sqrt{1 + \frac{C_{p1}}{C_{11}}}\sqrt{1 + \frac{C_{p2}}{C_{22}}}}\right) \cdot \left(\frac{2\sqrt{N - 1}}{\sqrt{(1 + \mu_1)(1 + \mu_2)}}\right) \tag{A1-16}$$

This expression for the gain-band figure of merit of a two-circuit, line-to-line amplifier is particularly useful for grounded-cathode pentodes and klystrons. It is the product of three factors. The first may be called the electronic figure of merit because it depends only upon electron stream parameters (ratio of transadmittance to mean capacitance of the electronic transducer T_e). The second is the degradation factor giving the effect of adding passive circuit capacitance both inside and outside the bulb to the active capacitance already present in the electronic transducer. The third factor, called the matching factor, depends only on the matching conditions and on the arbitrary definition of bandwidth. If the band is taken 6 db down (3 db for eacy circuit) and the tube input and output are matched, the third factor is unity.

In amplifiers using triodes and tetrodes in grid-return circuits, the Q of the input circuit is usually very much smaller than that of the output. Here it is appropriate to use the single-circuit limiting concept, with $Q_{1e} << Q_{2e}$. Here a figure of merit independent of bandwidth is obtained from the product of power gain and bandwidth:

$$|\Gamma_0|^2 B_N = \left(\frac{|Y_{21}|^2}{4\pi G_{11} C_{22}}\right)\left(\frac{1}{\left(1+\frac{G_{p1}}{G_{11}}\right)\left(1+\frac{C_{p2}}{C_{22}}\right)}\right)\left(\frac{8\mu_1\sqrt{N^2-1}}{(1+\mu_1)^2(1+\mu_2)}\right) \qquad \text{(A1-17)}$$

This expression for the gain-band figure of merit of a one-circuit, line-to-line amplifier is also the product of three factors. The first is again the intrinsic electronic figure of merit of the active transducer alone; the second is the degradation produced by the addition of passive circuit capacitance to the output and circuit loss to the input; the third is a band-definition matching factor which is unity when the band is taken 3 db down and the tube is matched.

In the application of the figures of merit, the third factors are usually omitted, since they depend only on the matching conditions and on the particular definitions of bandwidth used.

Power-Band Figure of Merit

In the problem of power output amplifier stages, the design specification of greatest importance is the bandwidth over which a certain power output can be obtained with a specified maximum of distortion. Of the many methods of specifying distortion, one which is particularly useful for microwave systems is known as the "compression". If the power gain is plotted in decibels as a function of the power output, as shown in Fig.

A1-3, it will normally be constant for low power levels (for which the device is essentially linear) and equal to the low level power gain $|\Gamma|^2$. However, at some higher power level non-linearities appear in some or all of the various short-circuit admittances, usually causing the power gain to decrease below the small-signal value by an amount called the compression, C. If P_0/P_i be power gain for any power output and $|\Gamma|^2$

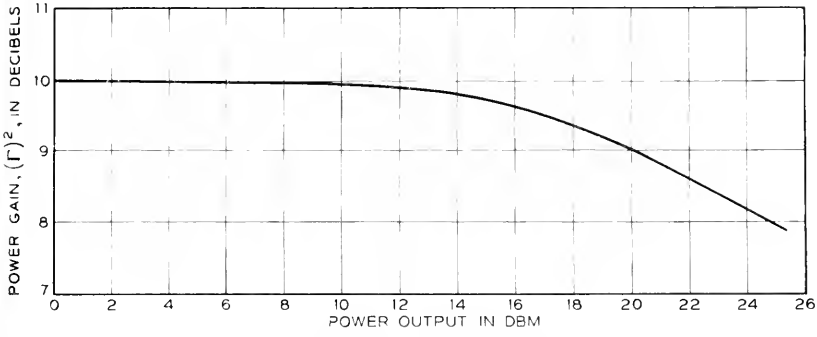

Fig. A1-3.—Typical gain variation with power output.

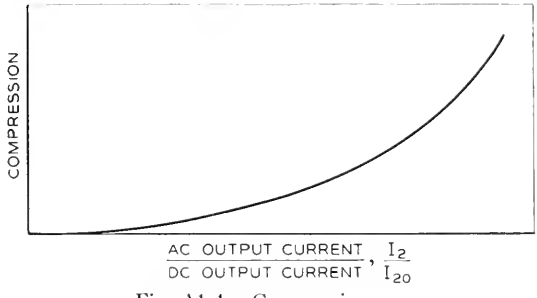

Fig. A1-4.—Compression vs. Alternating current in output / Direct current in output

the small-signal power gain, the compression C is defined in decibels as follows:

$$C = 10 \log_{10} |\Gamma|^2 - 10 \log_{10} P_0/P_i \qquad \text{(A1-18)}$$

$$= 10 \log_{10} \frac{|\Gamma|^2 P_i}{P_0}$$

Naturally, the compression depends upon how hard the tube is driven. It is therefore a function of the amount of drive, which may be conveniently expressed in terms of the ratio of the alternating output current to the operating direct current, as in Fig. A1-4.

The power output depends on operating parameters thus:

$$P_0 = I_2^2 \frac{G_2}{(G_{22} + G_2)^2} \tag{A1-19}$$

As the output power level is continually raised, more and more current is required to drive the load, until finally the non-linear distortion limit is reached. The maximum output current is therefore limited to a certain proportion of the direct current I_{20}, thus:

$$I_{2m} = I_{20} \cdot F(C)\ F_2(\theta_2) \tag{A1-20}$$

where $F(C)$ shows the dependence upon the compression C and will naturally be the larger, the more the allowable compression. $F_2(\theta_2)$ indicates a dependence upon output transit angle; it is the output gap coupling coefficient.

The power output depends also upon the output circuit conductance G_2 and can be greater if G_2 is smaller. However, a smaller G_2 implies a smaller bandwidth. It results that the power is inversely proportional to the bandwidth of the output circuit, or in other words, the product of power output by the bandwidth of the output circuit is a constant—a figure of merit of the tube. As in the case of the gain-band merit, this also can be broken up into factors:

$$P_0 \cdot B_N = \left(\frac{I_{20}^2 F^2(C) F_2^2(\theta_2)}{4\pi C_{22}}\right)\left(\frac{1}{1 + C_{p2}/C_{22}}\right)\left(\frac{2\sqrt{N^2 - 1}}{1 + \mu_2}\right) \tag{A1-21}$$

This expression for the power-band figure of merit is the product of three factors. The first is the intrinsic figure of merit of the active transducer alone; the second is the degradation caused by the addition of passive circuit capacitance to the output circuit; the third is a band definition—matching factor which is unity when the output is matched and the band of the output circuit is taken 3 db down.

The power-band computation does not depend upon the input circuit. Variations in the latter affect the gain of the amplifier, but not its overload point. Accordingly in the power band formula only properties of the tube and its output circuit appear. When feedback has to be considered, then the input circuit also affects the power, and the analysis becomes more complicated.

We have now three figures of merit: namely, two gain-band products applying to different kinds of amplifiers, and one power-band product. They relate the performance of an amplifier to certain internal parameters. For wide band service, the tube design should make the appropriate figure of merit as large as practicable.

It should be understood that many other factors may have a bearing on amplifier design, such as power consumption, noise performance or amount of feedback. Where such factors are important, they too must be considered, and frequently appropriate merit figures like plate efficiency or noise figure are useful.

APPENDIX 2

The Circuit Capacitance Degradation Factor

The capacitance degradation factor $C_{22}/(C_{22} + C_{p2})$ which applies to both gain-band and power-band products, can be calculated approximately

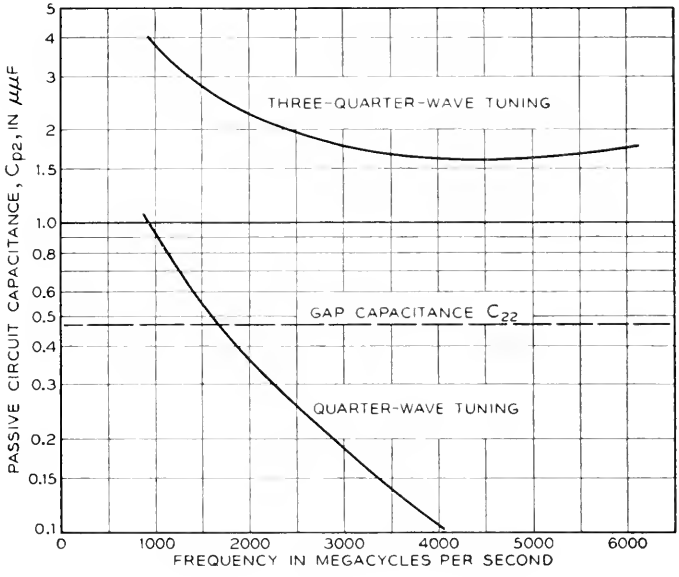

Fig. A2-1.—Passive circuit capacitance C_{p2}.

as shown below. As the frequency is varied, this factor changes by considerable amounts for the 1553 tube; accordingly, both figures of merit vary with frequency, and design control has been exercised to produce maximum merit around 4000 megacycles.

The capacitance degradation factor is just the proportion which the active tube capacitance bears to the total capacitance of tube and circuit, and would therefore have a maximum of unity if the circuit passive capacitance were made zero. For the 1553, we may begin by assuming that the plate circuit is to be tuned by a resonant coaxial line. As the frequency is lowered the effective capacitance will be increased, since the line must be lengthened; its variation is shown in Fig. A2-1.

The calculation is based on the following assumptions (Fig. A2-2):

1. The output cavity has inner diameter .180″, outer .850″, consequently a characteristic admittance G_0 :

$$G_0 = 7250/\log \frac{d_2}{d_1} = 10{,}710 \text{ micromhos} \tag{A2-1}$$

2. The gap capacitance is that of a parallel plate condenser of .180″ diameter and .012″ spacing, namely

$$C_{22} = \epsilon_0 A/d = 0.477 \ \mu\mu f \tag{A2-2}$$

3. The effect of the glass vacuum envelope is neglected for simplicity.

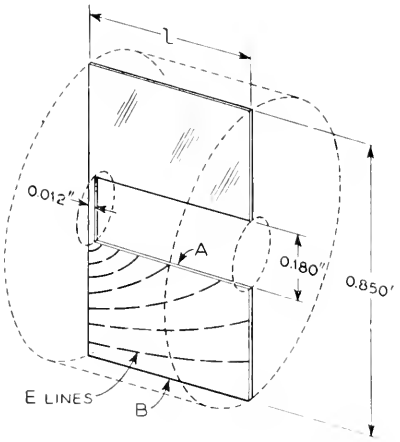

Fig. A2-2.—Output cavity dimensions. A, B are concentric cylindrical portions. Actual lines of electric force are partly dotted into sketch.

Consequently the length l of the line is given by the well-known tuning relation

$$\omega C_{22} = G_0 \cot \theta = G_0 \cot \frac{\omega l}{c} \tag{A2-3}$$

The distributed capacitance of the line is determined from the formulas (3) of the text, which in this case reduces to the following:

$$\omega C_{p2} = \frac{G_0 \theta}{2}\left(1 + \frac{\omega^2 C_{22}^2}{G_0^2}\right) - \frac{\omega C_{22}}{2} \tag{A2-4}$$

The cavity distributed capacitance is thus comparatively easy to calculate at high frequencies because of the simplicity of the geometry. At low frequencies the computation of the distributed capacity of a coil is no

different in principle, but would be harder to carry out in practice because of the helical geometry. The value can of course in any case be found by measurement of the tuning admittance as a function of frequency. From these equations the circuit degradation factor can be calculated, and is shown in Fig. A2-3 as a function of frequency.

The accuracy of the coaxial line assumptions decreases as the cavity becomes shorter. For 4000 and 6000 megacycles, since the length of the cavity is less than its diameter, it would be more nearly correct to regard it as a radial transmission line loaded by the inductive "nose" in the

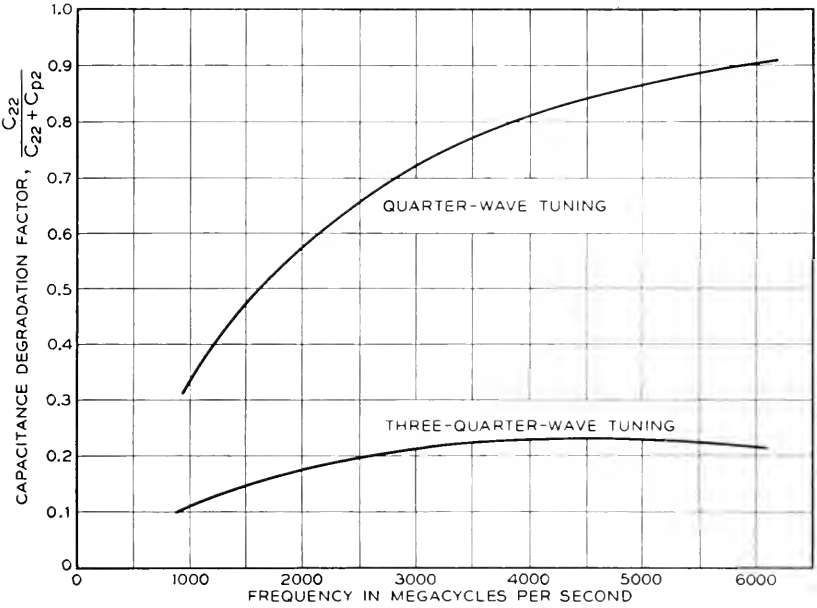

Fig. A2-3.—Capacitance degradation factor, $\frac{C_{22} + C_{p2}}{C_{22}}$.

center. The admittance of such a cavity can be calculated[12] or measured; but the additional precision hardly warrants the effort in the present case.

The capacitance degradation factor at 4000 megacycles is indicated from Fig. A2-3 as .81, or only 0.9 db less than the intrinsic limit of unity if the passive capacitance were entirely negligible compared to the active 0.5 $\mu\mu f$. This indication is somewhat optimistic, as appears from Fig. A2-2. The coaxial line formulas assume that the capacitance corresponds to a radial electric field between concentric cylinders A and B. This capacitance is found to be quite small (.11 $\mu\mu f$ at 4000 Mc.). The actual lines of

[12] S. Ramo and J. R. Whinnery, "Fields and Waves in Modern Radio," N. Y., Wiley, 1944.

force, dotted in the figure, clearly correspond to a somewhat larger capacitance, especially when the length of the cavity is smaller than its diameter; but this larger capacitance is probably still less than the active capacitance C_{22}.

In so far as the gain-band product depends on the circuit capacitance degradation factor (Fig. A2-3), the curve is probably fairly accurate up to 2000 megacycles and somewhat optimistic for higher frequencies where the coaxial line predictions are evidently too small.

Above 5000 megacycles the quarter-wave tuning cannot be used for the 1553 tube since the glass would interfere with the tuning plunger. A glance at Fig. A2-3 shows that moving the plunger back a half-wave to the next node involves a drastic loss in gain-band product—a factor of four at 6000 megacycles—because of the great increase in circuit passive capacitance. Redesign of the tube for good figure of merit at 6000 megacycles would therefore require the use of first-node tuning. A reduction in outer diameter would be necessary, and the use of an internal pre-tuned cavity might also be indicated.

A New Microwave Triode: Its Performance as a Modulator and as an Amplifier

By A. E. BOWEN* and W. W. MUMFORD

(*Manuscript Received Mar. 20, 1950*)

This paper describes a microwave circuit designed for use with the 1553-416A close-spaced triode at 4000 m.c. It presents data on tubes used as amplifiers and modulators and concludes with the results obtained in a multistage amplifier having 90 db gain.

Introduction

MICROWAVE repeaters are of two general types: those that provide amplification at the base-band or video frequency and those that amplify at some radio frequency. Of the latter there are two types: those that involve no change in frequency and those that do involve a change in frequency, that is, the radiated frequency is different from the received frequency. The Boston-New York link[1] is of this last type as is also the New York-Chicago link. This paper deals chiefly with a discussion of the application of the close-spaced triode[2] in a repeater of the type to be used between New York and Chicago.

A block diagram of this type of repeater appears in Fig. 1. The received signal comes in at a frequency of, say, 3970 mc. It is converted to some intermediate frequency, say 65 mc, in the first converter which is associated with a beating oscillator operating at a frequency of 3905 mc. After amplification at 65 mc it is converted in the modulator back to another microwave frequency 40 mc lower than the received signal and then it is amplified by the r.f. amplifier at 3930 mc and transmitted over the antenna pointed toward the next repeater station. Our attention will be focussed upon the performance of the close-spaced triode in the transmitting modulator and in the r.f. power amplifier in this type of repeater.

The close-spaced triode was assigned the code number 1553 during its experimental stage of development and, with subsequent mechanical improvements, it became the 416A. Some of the data reported herein were taken on one type, and some on the other; references to both the 1553 and 416A tubes will be noted throughout the text. The difference in electrical performance was not significant.

An early experimental circuit for the 1553 type tube will be described

* Deceased.

in detail and the performance as amplifier and modulator will be presented. Measurements of noise figure will be included with a discussion of the performance of multistage amplifiers.

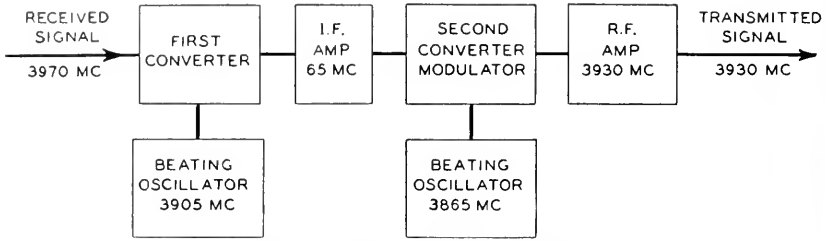

Fig. 1.—Typical microwave repeater.

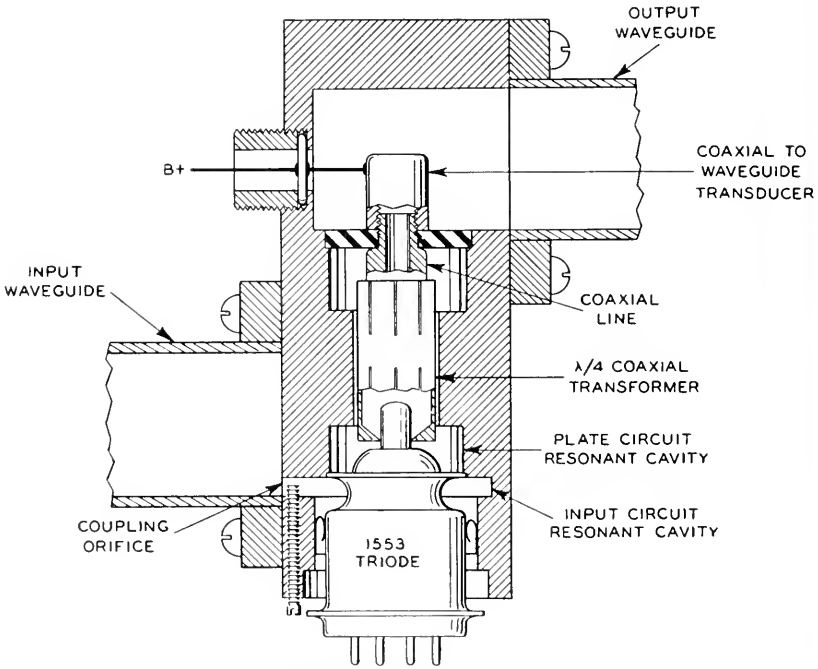

Fig. 2.—Microwave circuit for 1553 triode.

The Microwave Circuit

The experimental circuit which has to date met with greatest favor consists of cavities coupled to input and output waveguides, as shown in Fig. 2. The grid, of course, is grounded directly to the cavity walls and separates the input cavity from the output cavity. An iris with its orifice

couples the input waveguide to the input cavity and is tuned by a small trimming screw across its opening. The metal shell of the base of the tube makes contact to the input cavity through spring-contact fingers around its circumference and forms a part of the input cavity. The cathode and its by-pass condenser, located within the envelope, complete the input circuit cavity. The heater and cathode leads, brought out through eyelets in the base of the tube, are isolated from the microwave

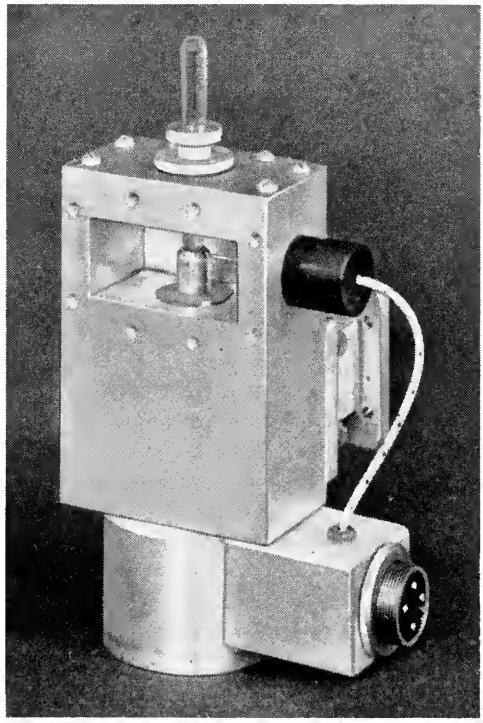

Fig. 3.—Model eleven microwave circuit for the close-spaced triode looking into the output waveguide.

energy in the input cavity by means of the internal by-pass condenser. When the tube is used as a modulator, this by-pass condenser acts as a portion of the network through which the intermediate frequency signal power is fed onto the cathode.

The output circuit cavity is coupled to the output waveguide through a coaxial transformer, a coaxial line and a wide-band coaxial-to-waveguide transducer. The output cavity is bounded by the grid, the coaxial line outer conductor, the radial face of the quarter-wave coaxial transformer

and the sealed-in plate lead of the tube. The plate impedance of the tube is transformed by the resonant cavity to a very low resistance (a fraction of an ohm) on the plate lead just outside the glass seal. The quarter-wave coaxial transformer serves to match this low impedance to the surge impedance of the coaxial line (45 ohms). Coarse tuning is accomplished by moving the slug of the outer conductor; fine tuning by moving the inner conductor. The coaxial line is supported at its end by a dielectric washer. Plate voltage is applied to the tube through a high impedance quarter-wave wire brought out to the low impedance probe through the side wall of the waveguide. Both the modulator and the amplifier used this type of circuit, which we call model eleven.

Fine tuning of the plate cavity is obtained by sliding the inner conductor of the coaxial transformer up and down on the plate lead. This movement is derived through a low-loss plastic screwdriver inserted through the hollow probe transducer; the driving mechanism is housed inside the inner conductor of the transformer, thus isolating the mechanical design problem from the electrical design problem effectively. The hollow stud at the top of the structure serves two purposes: screwing it into the waveguide introduces a variable capacitive discontinuity which serves to improve the match between the cavity and the waveguide. The length of the hollow plug provides a length of waveguide beyond cutoff which keeps the r.f. energy from leaking out through the plastic tuning screwdriver.

The heater and cathode leads from the tube are housed in a cylindrical metal can and are brought out through by-pass condensers to a standard connector. The photograph, Fig. 3, illustrates these features.

The input face of the circuit is illustrated in Fig. 4. The long narrow slot near the base of the rectangular block is the iris opening which couples the input waveguide to the cathode-grid cavity. The single tuning screw provided at the input iris is not adequate to match all of the tubes over the whole frequency band of 500 megacycles; an auxiliary tuner shown at the right of the circuit provides the necessary flexibility. This tuner, described by Mr. C. F. Edwards of the Bell Telephone Laboratories[3], is, in effect, two variable shunt tuned circuits about an eighth of a wavelength apart in the waveguide. Each variable tuned circuit is made up of a fixed inductive post (located off center in the waveguide) and a variable capacitive screw. It is capable of tuning out a mismatch corresponding to four db standing wave ratio of any phase.

As shown in Fig. 5, the tube slides into the bottom of the circuit and the grid flange is soldered to the wall of the cavity with low melting point

solder.† The shell of the tube is grasped by the springy contacts around the bottom of the input cavity. Above the tube the plate lead projects into the cylindrical space which can be adjusted to the desired size by the quarter wave slug seen to the right of the circuit. This makes contact to the walls of the outer cylinder by spring fingers on each end. Contact to the plate lead is then made through the movable slotted inner conductor, seen on the extreme right of Fig. 5.

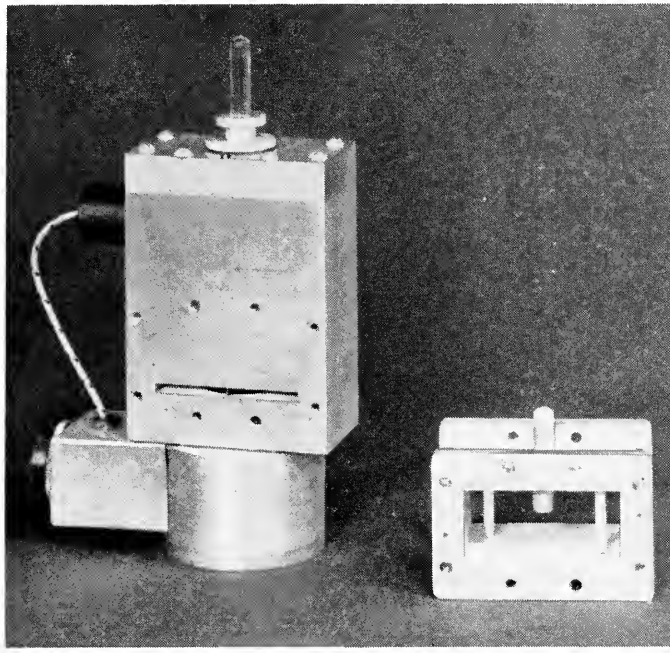

Fig. 4.—The input face of the circuit.

Figure 6 gives an exploded view of the details of the circuit, showing the simplicity of the construction which permits easy assembly. The guide pin which serves to keep the inner conductor of the transformer from rotating as it slides up and down on the plate lead during the tuning process can be seen on the third detail to the right of the main block. Also there is provision for external resistive loading to be introduced into the plate cavity through the small square holes in each side of the block. A screw mechanism adjusts the penetration of the loading resistive strip into the

† The early experimental tubes were soldered into the circuits. Chiefly through the efforts of Mr. C. Maggs and Mr. L. F. Moose, of B. T. L., who undertook the development of the tube for production by the Western Electric Company, the present 416A tubes come with a threaded grid flange to facilitate replacement.

plate cavity to provide for a limited adjustment of the bandwidth of the circuit. These are not always used, however, and most of the data to be presented here are for the condition of no external resistive loading.

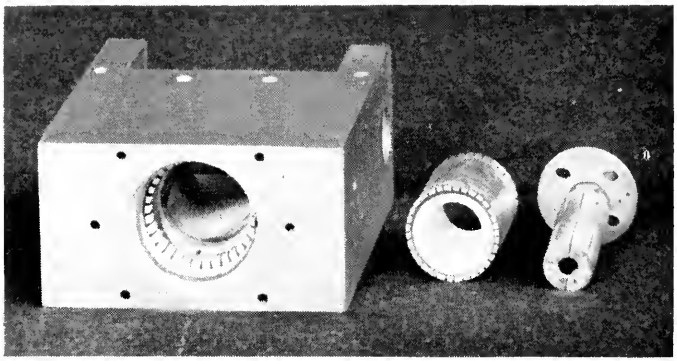

Fig. 5.—Bottom view of circuit.

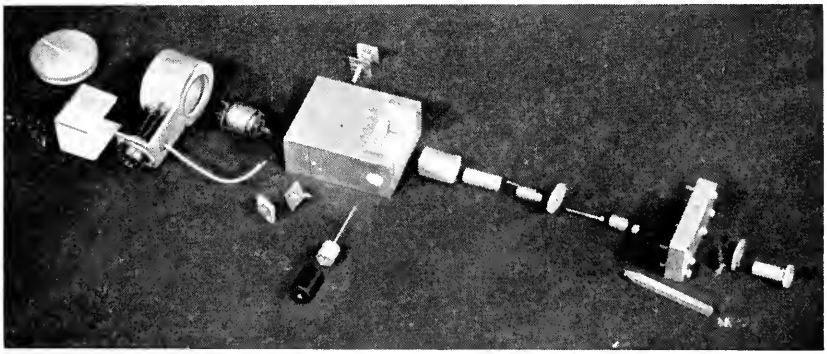

Fig. 6.—Exploded view of details.

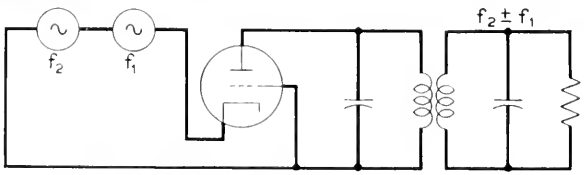

Fig. 7.—Elementary grounded grid converter schematic.

Modulator

The grounded grid transmitting converter shown schematically in Fig. 7 includes the two generators, a microwave beating oscillator, f_2, and an intermediate frequency signal, f_1, which impress voltages on the cathode,

the grid itself being grounded. The output circuit in the plate is tuned to the sum or difference frequency, $f_2 \pm f_1$.

By-pass condensers, traps and filters for other frequencies present in the modulator must be considered. Besides the beating oscillator and the signal, their sum and difference frequencies appear in both the input circuit and the output circuit and of course bias voltage on the cathode and plate voltage on the plate must be applied. Some of the traps and by-pass condensers which influence the converter performance are indicated in Fig. 8. It is obvious that microwave energy should be kept from flowing into the i.f. signal circuit and vice versa if the highest conversion gains are to be obtained. Both of these conditions are easily achieved. It is not so readily apparent that the components of the wanted and the unwanted sidebands present in the input circuit must be handled

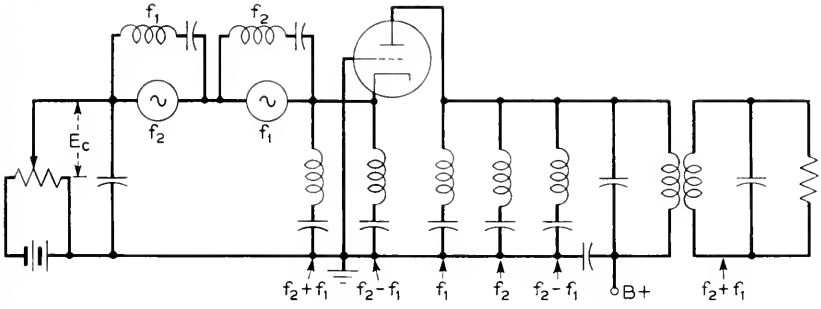

Fig. 8.—Schematic diagram of converter with traps and filters for fundamental frequencies of signal, f_1, beating oscillator, f_2, and sidebands, $f_2 \pm f_1$.

properly. Of these two, the more important is the wanted sideband and the next figure illustrates just how necessary it is to treat it properly.

The simplest way to keep the wanted sideband component of the input circuit from being absorbed by the beating oscillator branch is to reflect the energy back into the converter by means of a reflection filter. This reflected energy arrives back at the tube and may conspire to reduce the conversion gain of the modulator if the phase is wrong. The phase depends upon the spacing along the waveguide between the tube and the filter and Fig. 9 illustrates how badly the gain is affected when the wrong spacing is used. Data for two different tubes are given which indicate that the correct spacing for one tube may be incorrect for another. It should be pointed out, however, that these two tubes were early experimental models and that production tubes behave more consistently.

The i.f. impedance of the modulator is also affected by the filter spacing for the wanted sideband on the input. This effect can be utilized to

vary the i.f. impedance by small amounts to achieve a better i.f. match, since the proper spacing for best gain is not a critically exact dimension. That is to say, there is a fairly large range of spacings which give good performance as far as conversion gain is concerned so that, as long as the critical distance which gives poor gain is avoided, the i.f. impedance can be adjusted by varying the spacing of the input filter.*

It is important that the i.f. impedance of the modulator be adjusted to match the impedance of the i.f. amplifier which drives it, since any mismatch would cause a degradation of the system performance. In the design of the matching transformer the inductance of the leads, the capacity of the tube and by-pass condenser and the resistance of the elec-

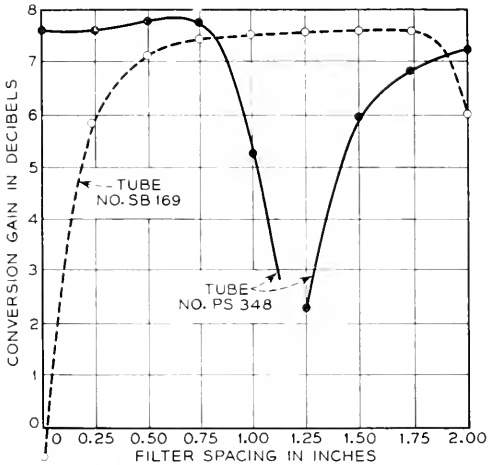

Fig. 9.—Data showing the effect of the spacing of a rejection filter for the wanted sideband in the input circuit.

tron stream were measured at the base of the tube. A broad-band transformer was designed and the inductances were thrown into an equivalent T network, thereby utilizing the lead inductance inside the tube as a part of the transformer, absorbing it in the L_2-M branch as indicated in Fig. 10. In several experimental tubes the lead inductance was .04μH. The impedance match obtained with such a transformer gave less than two db SWR over a band from 55 to 75 mc with the loop at the cusp on the reflection coefficient chart characteristic of slightly over-coupled tuned transformers as shown in Fig. 11.

The broadband matching of the output circuit of the modulator required a different technique. Not only is this filter called upon to provide a broad-band impedance match, but also it should provide dis-

* The spacing of the input filter also affects the plate impedance in a complicated way.

crimination against the other microwave-frequency components present in the modulator output circuit; consideration of the beating oscillator and

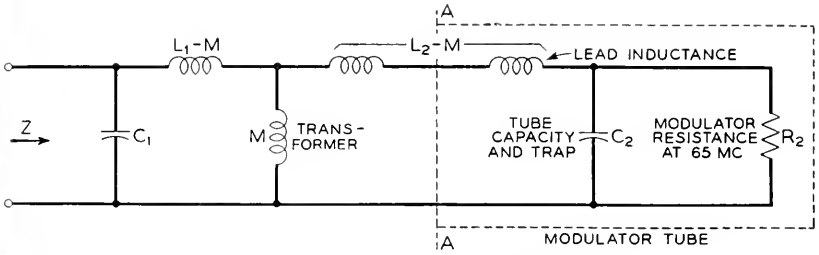

Fig. 10.—Equivalent circuit of modulator at I. F.

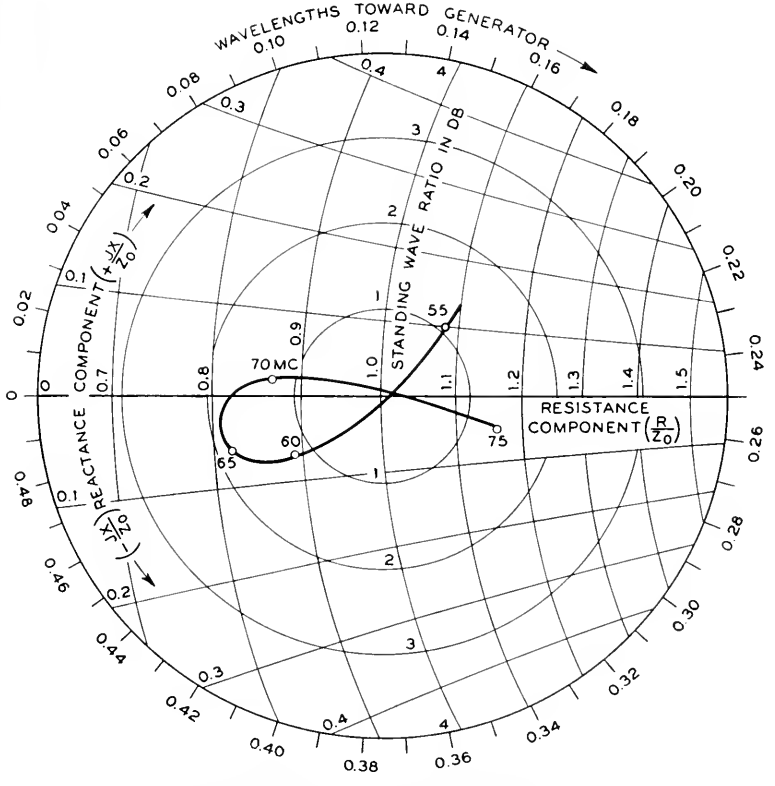

Fig. 11.—Modulator I. F. impedance with transformer.

both sidebands is necessary. The variables at our disposal are the bandwidth of the modulator output circuit and the number of cavity resonators which follow it. The desired quantities are the specified transmission

bandwidth and the attenuation required at the beating oscillator frequency. With two equations and two unknowns, the maximally-flat filter theory was applied to the circuit shown schematically in Fig. 12.[4] This indicated that an output circuit bandwidth of 84 mc (to the three db loss points), associated with two external resonant branches having bandwidths of 42 and 84 mc respectively, were needed to obtain a 20 mc flat band with 30 db suppression of the beating oscillator.

Such cavities were designed and attached to the output of a modulator whose bandwidth had been adjusted by means of small resistive strips.

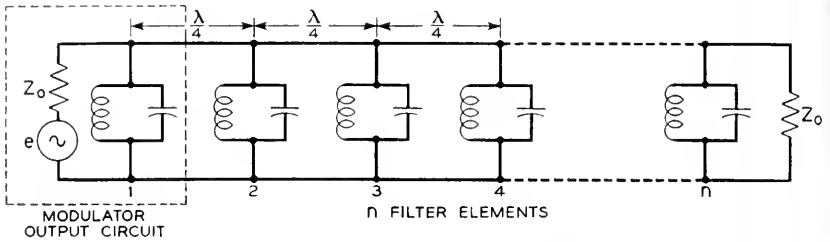

Fig. 12.—Sideband filter in waveguide.

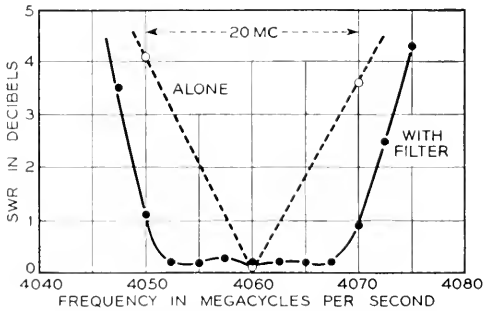

Fig. 13.—Output circuit impedance match.

The resulting impedance match gave a standing wave ratio of less than one db over a 20 mc band (the plate circuit alone without the filter was only about 5 mc wide to corresponding points) as shown in Fig. 13, and the beating oscillator power at the output of the filter was less than one tenth of a milliwatt, corresponding to 33 db discrimination.

The requirements and specifications for this particular experimental model do not necessarily reflect out present thoughts upon the requirements for any particular microwave radio relay system; they are presented here in some detail to indicate how certain specifications can be met, rather than to express what those specifications should be.

Other factors which influence the performance of the 416A modulator

are the plate voltage and the beating oscillator drive. The beating oscillator power affects the low level gain only slightly but has quite an effect on the gain at high power levels, that is, when the output power becomes comparable with the beating oscillator power. It is seen in Fig. 14 that compression becomes noticeable when the output power approaches within ten db of the beating oscillator driving power.

Varying the plate voltage on the modulator from 150V to 300V had little effect upon the conversion gain at low levels, but more power output

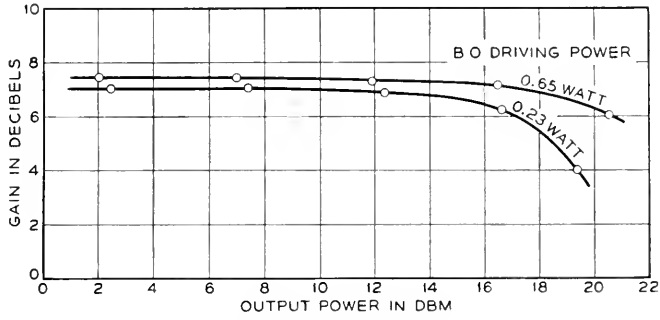

Fig. 14.—Modulator compression data for tube ✱PS62.

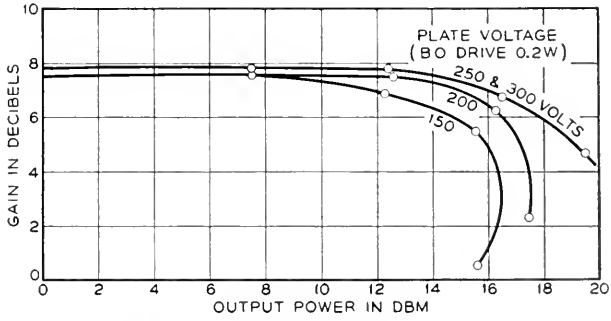

Fig. 15.—Modulator compression data for tube ✱PS348.

was obtained at the higher voltages. At 15 dbm power output, very little difference between 200V and 300V was observed, but at 150 V the gain was down two db, as shown in Fig. 15.

Fig. 16 shows the compression data for seven early experimental tubes, used as modulators with 200V on the plate, 14 ma cathode current, and 200 mw of beating oscillator drive. Half of these tubes had over seven db low level gain and only slight compression at power output levels of 13 dbm. The two poorest tubes would probably have been rejected before shipment, according to present standards of production. Each of the seven tubes was matched in impedance on the r.f. and i.f. inputs and also

on the r.f. output. The curves represent unloaded gain; no external loading was added to increase the bandwidth.

The performance of the close-spaced triode when used as a modulator appears to be superior in some respects to that of the silicon crystal modulators which are used in the New York-Boston microwave relay system.[1]

Single tubes had from 5 to 9 db gain compared with from 8 to 11 db loss for the crystals for corresponding power outputs. To get this performance the beating oscillator drive was only 200 milliwatts, compared with about 700 milliwatts for the crystal modulator. This reduction in r.f. power requirements means considerable simplification in a repeater.

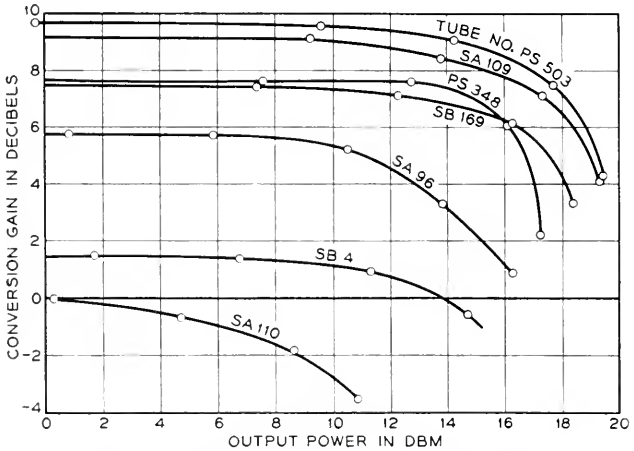

Fig. 16.—Compression data on seven 1553 triodes.
Plate voltage 200V
Plate current 14 Ma
B.O. Power 200 MW
Matched inputs and output

To offset this, the tube requires power supplies which are not necessary for the crystals, but low voltage power supplies should be cheap. The bandwidth of the tube modulator, 60 to 80 mc is less than the very wide (500 mc) band of the crystal modulator but it is comparable with the band width of the extra i.f. stages needed to drive the crystal modulator. The life of the tubes, although very little data are available as yet, will probably be less than the practically indefinite life of the silicon point contact modulators.

Amplifier

The performance of the close-spaced triode as an amplifier can best be described by referring to its impedance match, gain, transmission bandwidth and compression.

In some of the experimental tubes, bandwidths to the half power points of 21 mc to 250 mc have been measured. Typical of one of the better tubes, though not the best one, are the data contained in Fig. 17. The bandwidth of the input circuit is about twice that of the output circuit, and the SWR slumps outside the band on the low frequency side. The output impedance is more regular, exhibiting the familiar standing wave

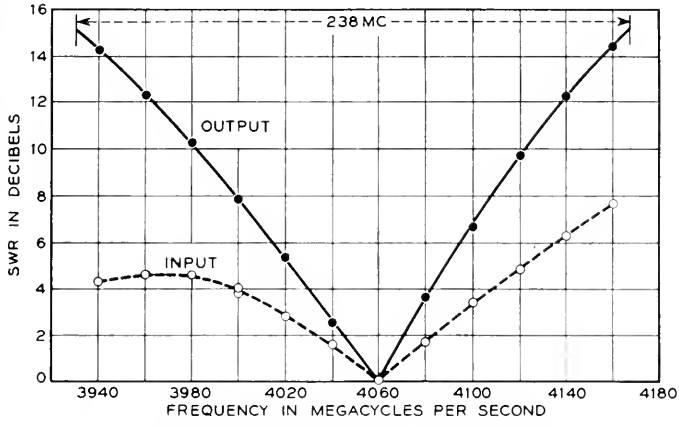

Fig. 17.—Input and output standing wave ratio versus frequency.

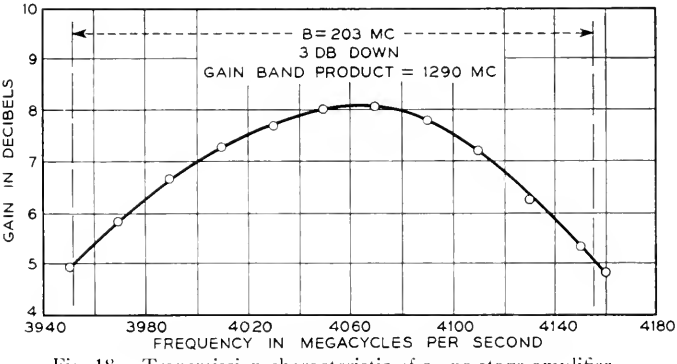

Fig. 18.—Transmission characteristic of a one-stage amplifier.

ratio of a simple single tuned resonant circuit. When the output impedance is plotted on the Smith reflection coefficient chart, the circle which results is also similar to that of a single tuned circuit. This is desirable since it then becomes a simple matter to incorporate the plate circuit in a maximally-flat filter of as many resonant branches as are needed, in the same way that the modulator output circuit was treated.

The transmission bandwidth for this single stage amplifier was 203 mc to the half power points, as shown in Fig. 18. This, with a gain of 8.05 db

at midband, gave a gain-band product of 1290 mc. The bandwidth of 203 mc was considerably greater than the average for these tubes. Similar results on 35 experimental tubes yielded the following averages: Low-level gain 10 db; Bandwidth 103 mc; Gain-band product 916 mc. The 416-A tubes produced by Western Electric Company exhibit comparable averages with much less spread; for example, a recent sample of 138 tubes had average values and standard deviations as follows:

TABLE I
GAIN AND BANDWIDTH OF 138 W. E. CO. 416-A TRIODES

	Average	St'd. Dev.
Low-level gain	9.9 db	1.1 db
Bandwidth	110 mc	9 mc
Gain-band product	1080 mc	350 mc

It is indeed gratifying to realize that such a remarkable tube can be produced with such uniformity.

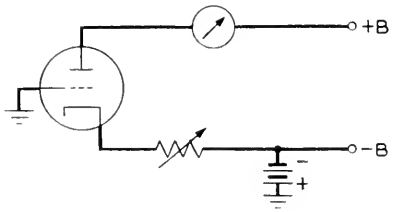

Fig. 19.—Stabilizer circuit.

In operating these tubes, it has been found that small variations in gain due to power line fluctuations and due to other disturbing influences can be minimized by using a stabilizing bias network which provides a large amount of negative feedback for the dc. path. This circuit is similar to one proposed by Mr. S. E. Miller of the Bell Telephone Laboratories for use in coxial repeaters which also use high transconductance tubes. In this circuit, shown schematically in Fig. 19, a few volts negative are applied to the cathode through a suitable dropping resistor. In the absence of plate voltage, the grid draws current, being positive with respect to the cathode. When plate voltage is applied, the drop in the cathode resistor tends to bring the cathode nearer ground potential until a stable voltage is reached. The resistor is set to a value which allows the desired cathode current to flow and subsequent variations in g_m or plate voltage then have little effect on the total cathode current.

Maintaining the cathode current constant does have an appreciable effect on the gain of the tube when operating at high output levels. This is characterized by a decrease in gain as the driving power is increased.

Fig. 20 illustrates this point. The low-level gain of this tube was 12.3 db but when the tube was driven so as to have an output power of 400 mw the gain was only about 3 db. At this point, retuning the circuit to rematch the tube at the high output level increased the gain to about 5 db. Now, returning to low level, the gain was only 10 db. Presumably in between these two points, 5 db at 500 mw output and 12.3 db at less than one milliwatt output, the performance could have been better than either of these two curves shows, i.e., the performance could have been improved by rematching at each intermediate power level.

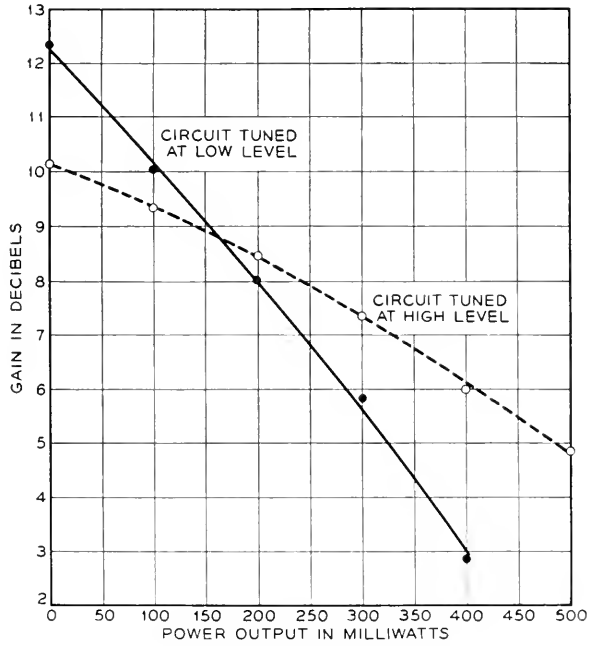

Fig. 20.—"Compression" in a one-stage microwave amplifier $I_p = 30$ ma.

This tube is not representative of all of the tubes tested. It is rather poorer in the spread of the two curves than most. It was picked merely to illustrate that besides a drop in gain also a detuning effect takes place when the driving power is changed. In the example given here the cathode current was held at or near 30 ma by the stabilizing bias circuit.

Without the stabilizing circuit, these so-called "compression" curves would be quite different. For instance, if the bias were held constant, we should expect that the gain would not drop as fast as indicated here, since the plate current would rise as the drive was increased.

At any rate, in an F.M. system, we are not concerned with how much

"static compression" exists, but rather with how much gain can be realized without exceeding the dissipation ratings of the tubes.

With this in mind data were taken on 25 of the experimental tubes. In each case they were matched to the input and output waveguides and the cathode current was stabilized at 30 ma. After driving the tube to a high level of output power, the circuits were rematched and the resulting "compression" curves revealed the capabilities tabulated.

TABLE II
SUMMARY OF DATA ON 25 EXPERIMENTAL CLOSE-SPACED TRIODES

	Highest	Lowest	Average
Low level gain	12.3 db	3.8 db	7.8 db
Gain (500 mw output)	7.0 db	−8.0 db	1.82 db
Power Output (3 db gain)	950 mw	50 mw	455 mw

It can be seen from the table that we might expect to obtain a gain of 20 or 25 db with three or four stages with a power output of about 500 mw and a flat band of over 20 mc.

THREE STAGE AMPLIFIER

A three-stage amplifier with 24 db gain has been assembled using an earlier type of circuit and loop tested at low levels on the equipment of Messrs. A. C. Beck, N. J. Pierce and D. H. Ring.[5] This amplifier had a bandwidth of about 30 mc to the 1 db points and while it does not represent the best that can be done with the 416A tube, the results of the loop test are interesting.

The recirculating pulse test, or loop test, is performed on a repeater component to determine its ability to reproduce a pulse faithfully after repeated transmissions. The output of the amplifier is connected to its input through a long delay line and an adjustable attenuator. The overall gain of the loop thus formed is adjusted to unity or zero db so that an injected pulse will recirculate through the loop without attenuation but accumulating distortion with each round trip. After allowing the pulse to recirculate long enough the amplifier is blanked out or quenched and the recirculating pulse amplitude dies out, thus preparing the loop for the next injected pulse, when the process is repeated. With a pulse length of one microsecond and an overall delay of two microseconds, one hundred round trips occur in 0.2 milliseconds, thus allowing the process to be repeated at the rate of two or three thousand times per second. A cathode ray oscilloscope is used to examine the pulse shapes, and its sweep is synchronized to the injected pulse so that successive corresponding pulses are superposed, enabling the operator to examine the pulse after any

number of round trips or select individually the nth round trip for inspection.

Fig. 21(a) shows the complete cycle between successive injected pulses, and the individual pulses that follow cannot be resolved at this slow sweep speed. Fig. 21(b) shows the first 26 round trips resolved so that they are distinguishable. Figure 21(c) shows the first and second round trips

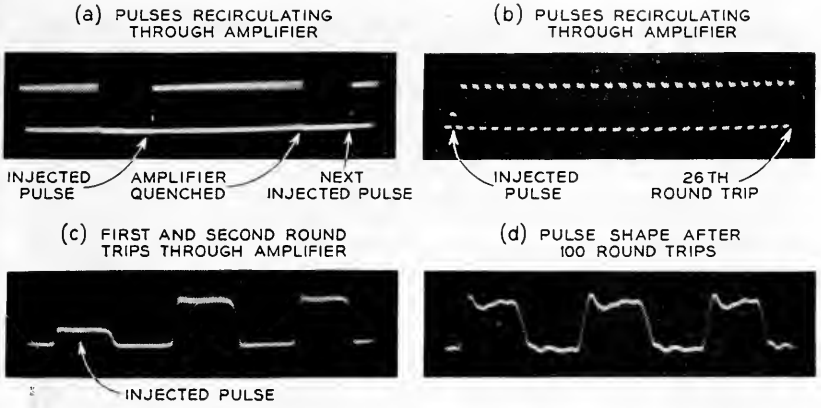

Fig. 21.—Recirculating pulse test patterns.

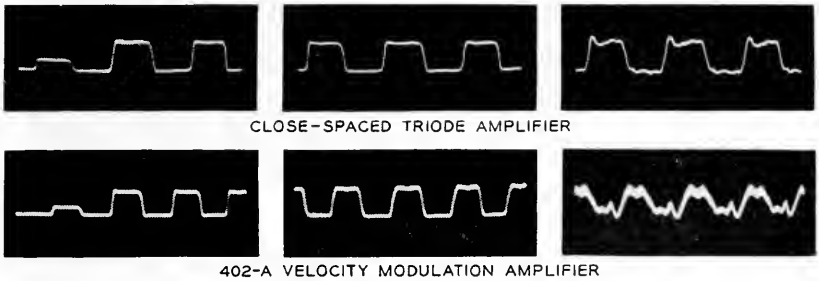

Fig. 22.—Recirculating pulse patterns showing 1st, 10th and 100th round trips for: Top: Close-spaced triode amplifier. Bottom: 402-A velocity modulation amplifier.

through the amplifier, with little or no distortion discernible. Fig. 21(d) gives, to the same scale as the preceding picture, the pulse shape after 100 round trips. A little overshoot and subsequent oscillation is now visible, although the whole pulse shape is still not too bad.

In Fig. 22, these results are compared with the results of a similar test performed on a four-stage, stagger tuned, stagger-damped amplifier using the 402 velocity variation amplifier tubes; the first, the tenth and the hundredth round trips are shown. Little or no distortion is seen at the

tenth round trip, but the superiority of the 416A amplifier is clearly shown in the hundredth round trip.

Both amplifiers were operating at low levels, and the pulse was an amplitude modulated one. Since these are not the conditions under which our microwave radio relay circuits operate, conclusions should not be drawn about how many repeater stations can now be put in tandem. The

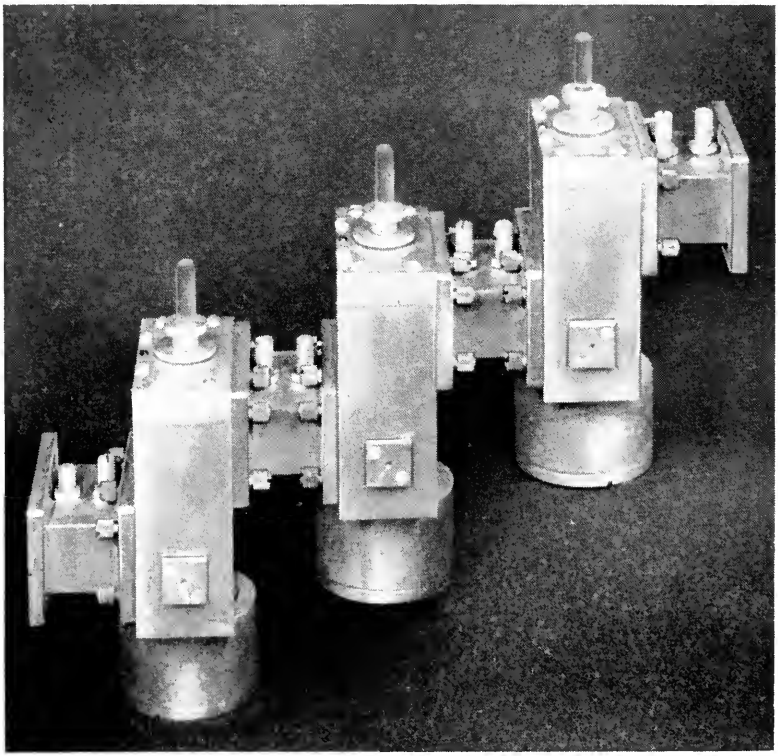

Fig. 23.—An assembled three-stage microwave amplifier.

test merely indicates that an improvement has been made, thus corroborating the evidence obtained by other tests.

Still further improvement has been made since loop testing the model ten amplifier. A three stage 416A amplifier (see Fig. 23) using model eleven circuits had comparable gain, 23 db, but a bandwidth of 50 mc to points 0.1 db down. These data again are for low level operation, but it is reasonable that half a watt might be expected from four such stages with comparable gain and slightly narrower bandwidth, surely 30 mc.

Noise Figure

In a forward looking program it is well to keep in mind other possibilities for this tube, such as use in a straight through type of repeater in which all of the amplification is obtained at microwave frequencies. In such an application the noise figure of the triode becomes one of its limitations, since the 416A must compete with the low noise figure of the silicon crystal converter which, for the New York-Boston circuit, is around 14 db. Data on thirty five early experimental and production 416A tubes gave an average value of 18.08 db at 4060 mc.* Each of the

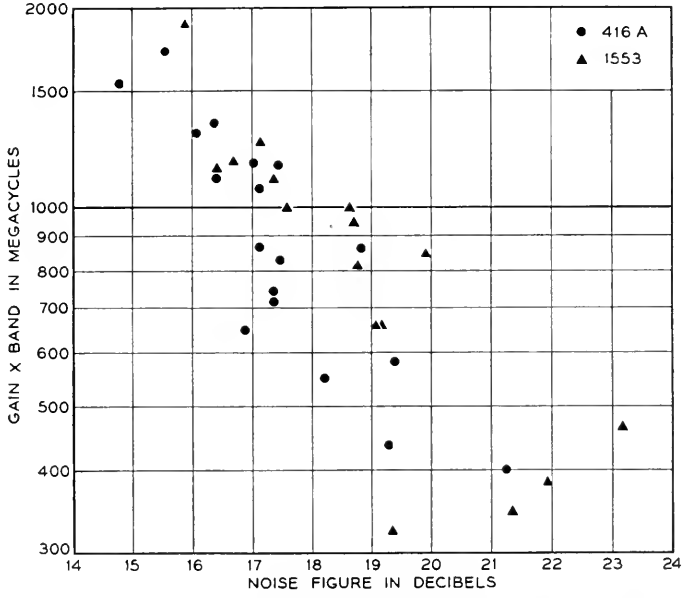

Fig. 24.—Noise figure vs gain-band product for close-spaced triode.

tubes was operated at 200 volts with 30 ma space current and was tuned so as to present matched impedances to the input and the output waveguides. The best of this batch had a noise figure of 14.79 db and the poorest 23.2 db. These measurements were made with a fluorescent light noise source.[6]

An interesting correlation between noise figure and gain-band product was uncovered during these tests, as can be seen in Fig. 24, which gives the noise figure in db on the abscissa and the gain-band product in mega-

* More recently, a sample of twelve production 416A tubes ranged from 13.5 to 16.2 db and averaged 15.06 db noise figure, with a standard deviation of 0.8 db.

cycles along the logarithmic ordinate. The points scatter between the extremes of 15 db noise figure for a gain-band product of 2000 megacycles to 23 db noise figure at 400 megacycles gain-band product. Extrapolating from these data, a noise figure of 10 db might be achieved if the gain-band product could be increased to 5500 mc. It is reasonable to expect that an improvement of this amount can be achieved if the resistance and return electron losses inside the tubes can be eliminated.[7]

We may use these data to determine the expected noise figure of a straight through amplifier, thus:

$$F = F_A + \frac{F_B - 1}{G_A} + \frac{F_C - 1}{G_A G_B} \cdots \tag{1}$$

If, for example, we assume that all stages are alike in noise figure and in gain, equation (1) approaches the expression, as the number of stages increases without limit:

$$F \underset{n\to\infty}{=} \frac{F_A G_A - 1}{G_A - 1} \tag{2}$$

Using an average value of 10 db gain per stage, the overall noise figure would be as follows:

(1) For $F_A = 30$ (best tube, 14.79 db)
$F = \frac{299}{9} = 33.2$ or 15.2 db
(2) For $F_a = 64$ (average tube, 18.08 db)
$F = \frac{639}{9} = 71$ or 18.5 db.

Straight-Through Amplifier

The actual performance of a ten-stage amplifier was about what should be expected from the considerations above. The best tube (10 log F = 14.79 db) was used in the first stage, and the next best tube in the second stage. The measured overall noise figure was 15.96 db. The overall gain was 90 db and the band was flat to 0.1 db for 44 mc. Such an amplifier with its associated power supply and individual control panels is shown in Fig. 25.

Conclusions

A circuit is described which lends itself readily to utilizing the 416A close-spaced triode as a modulator or a cascade amplifier for microwave repeaters operating at 4000 mc. Data are presented on early experimental models of the tube.

As modulators, single tubes had from 5 to 9 db gain with 10 to 20

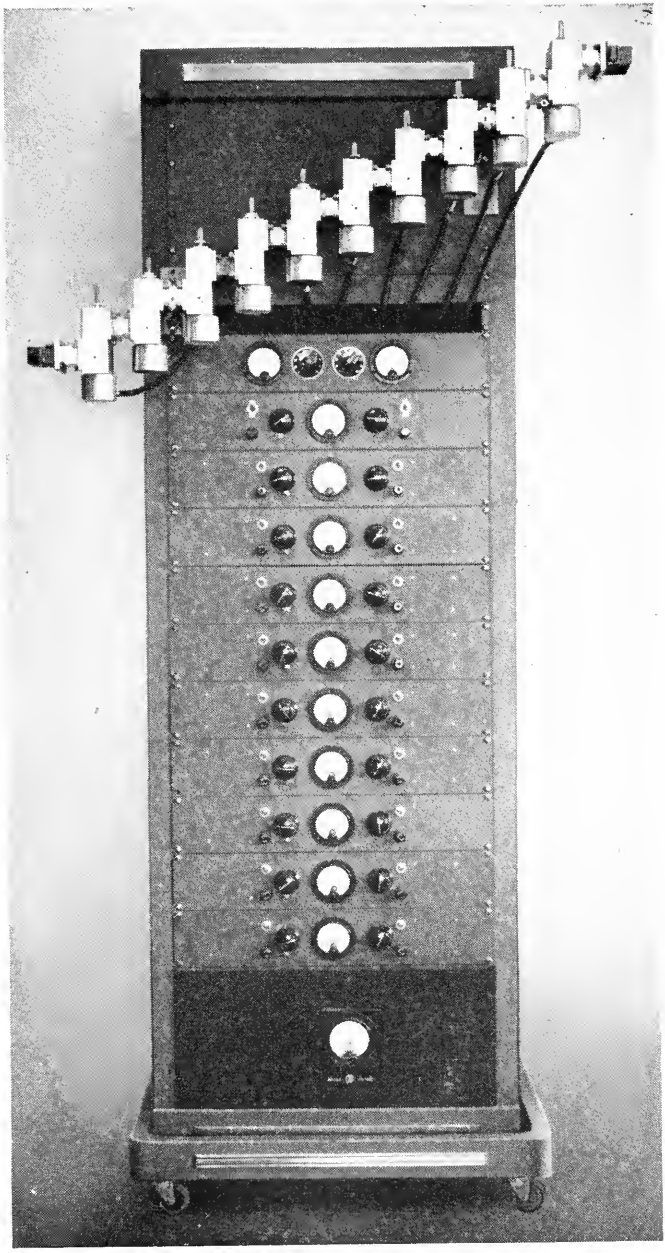

Fig. 25.—A ten-stage microwave amplifier operating at 4000 mc.

mw output when driven with 200 mw of beating oscillator power. A bandwidth of twenty megacycles was readily obtained.

As amplifiers at 4060 mc, the average gain of 60 tubes was 9 db, the average bandwidth of 34 tubes was 103 mc to the half power points, the average noise figure of 35 tubes was 18.08 db and the average power output (for 3 db gain) was 455 mw for 25 tubes. Operating the tubes in cascade produced an amplifier which had less distortion of pulse shape than an earlier amplifier which used the 402-A velocity variation tube. A ten-stage amplifier has been assembled and tested, yielding 90 db gain, a noise figure (with selected tubes) of 15.96 db and a bandwidth of 44 mc to the 0.1 db points.

These data are for early experimental models of the tube and it is likely that subsequent alterations may improve the performance in the production models.

Acknowledgments

The work described in this paper took place at the Holmdel Radio Research Laboratories. Mr. Bowen, with the able assistance of Mr. E. L. Chinnock, was active in pursuing the problems connected with the amplifier circuits. Mr. R. H. Brandt helped with the work in connection with the modulator circuits and many others were helpful in designing and constructing the circuits and facilities for testing.

References

1. "Microwave Repeater Research," H. T. Friis, *B. S. T. J.*, Vol. 27, pp. 183–246, April 1948.
2. "A Microwave Triode for Radio Relay," J. A. Morton, *Bell Labs. Record*, Vol. 27, #5, May 1949.
3. "Microwave Converters," C. F. Edwards, *Proc. I. R. E.*, Vol. 35, pp. 1181–1191, Nov. 1947.
4. "Maximally-Flat Filters in Wave Guide," W. W. Mumford, *B. S. T. J.*, Vol. 27, pp. 684–713, Oct. 1948.
5. "Testing Repeaters with Circulated Pulses," A. C. Beck and D. H. Ring, *Proc. I. R.E.*, Vol. 35, pp. 1226–1230, November 1947.
6. "A Broad-Band Microwave Noise Source," W. W. Mumford, *B. S. T. J.*, Vol. 28, pp. 608–618, October 1949.
7. "Electron Admittances of Parallel-Plane Electron Tubes at 4000 Megacycles," Sloan D. Robertson, *B. S. T. J.*, Vol. 28, pp. 619–646, October 1949.
8. "Design Factors of the Bell Telephone Laboratories 1553 Triode," J. A. Morton and R. M. Ryder, *B. S. T. J.*, Vol. 29, #4, pp. 496–530, Oct. 1950.

A Wide Range Microwave Sweeping Oscillator

By M. E. HINES

(*Manuscript Received July 24, 1950*)

1. Introduction

A SWEPT frequency oscillator is a useful laboratory tool for testing wide-band circuit components. It permits an oscillographic display of a frequency characteristic, avoiding much of the labor of point-by-point testing at discrete frequencies. There was a particular need in the Bell Telephone Laboratories for a sweeping oscillator to cover the communications band between 3700 and 4200 megacycles to facilitate the testing of components for radio relay repeaters.

This paper describes one type of oscillator designed to satisfy this need. It utilizes the BTL 1553 (or the Western Electric 416A) microwave triode. The tuning is accomplished mechanically so that the frequency varies continuously back and forth over the band at a low audio frequency rate. Continuous oscillations have been obtained over a 900 megacycle band from 3600 to 4500 megacycles.

2. Circuit Structure

Basically, the rf circuit consists of a tunable cavity for a grid-anode resonant circuit, a means for feedback to an untuned grid-cathode circuit, and a means for coupling the cavity to a waveguide output. The grid-anode cavity is the only sharply tuned circuit, and it was found that oscillations could be obtained over the entire band by changing the resonant frequency of that cavity alone. In this application, the electronic conductance between the grid and cathode is so high that this portion of the circuit has an inherent broad band such that separate tuning is unnecessary.

The necessity for continuous, rapid tuning virtually requires that there be no sliding contacts in the tuning mechanism. A type of cavity was chosen so that tuning could be accomplished by a simple variable capacitor of the non-contacting type. Reduced to its simplest elements, it consists of a short coaxial line, resonant in the half-wave mode. Actually the line is much shorter than a half wavelength because of excess capacitance at both ends. At one end is the capacitance of the grid-anode gap, and at the other end is the variable capacitor used for tuning.

The actual cavity is illustrated in Fig. 1. This is somewhat more complicated than a half wave line, but the mode of resonance is essentially the same. The variable capacitor utilizes a thin-walled copper cup which is movable vertically. This cup fits rather closely inside, and is coaxial with, a cylindrical hole in the main body of the cavity. It forms the center

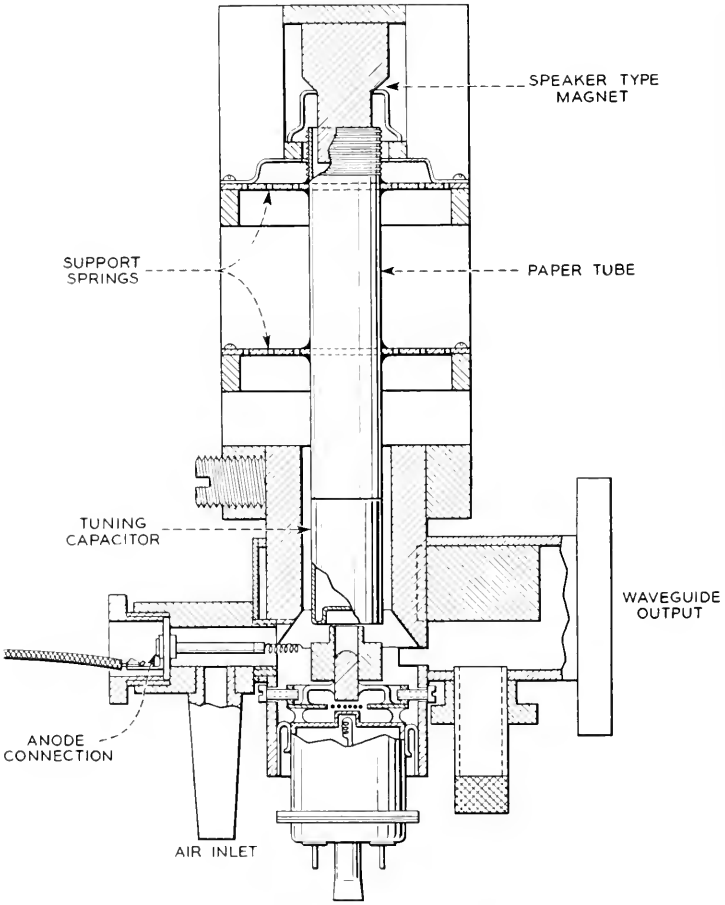

Fig. 1. Construction of the oscillator.

conductor of a low impedance coaxial line approximately one-fourth wavelength long, so that in this frequency range it is effectively short-circuited to the cavity wall. Vertical motion of the cup is therefore roughly equivalent to moving the end wall, thereby changing the capacitance between the wall and the center conductor of the main cavity. The reso-

nant frequency is lowest when the two surfaces are nearly in contact, and highest when the cup is fully extracted.

The recessed end of the cup fits over a protuberance on the center conductor when they are nearly in contact. This special shape was designed to give a reasonably straight curve of frequency vs. displacement. With planar surfaces, the frequency would change more rapidly with displacement at the low than at the high frequency end of the band.

The grid disk of the tube is separated from the wall of the cavity by a narrow annular space, and contact is made across the gap by a number of small screws. These screws act as an inductive reactance in series with the circulating currents of the resonant cavity. The voltage developed across this reactance is applied between the grid and the main envelope of the tube, and in this way energy is fed into the grid-cathode space to provide feedback.

The mechanical tuning device was adapted from an inexpensive permanent magnet loudspeaker of the type used in small home radios. The construction is shown in Figs. 1 and 5. The speaker cone was removed and the voice coil was attached to a thin-walled paper cylinder which supports the tuning cup inside the cavity. Two sheet fiber springs support the paper cylinder and maintain the axial alignment in the magnet and cavity. These springs are cut with a number of incomplete circular slits to reduce the stiffness for axial motion. With the voice coil actuated from a small filament transformer, peak to peak motion $\frac{1}{4}$ of inch is obtainable.

The heater and cathode connections are made at the base of the tube which protrudes from the cavity. The grid is internally connected to the main body of the cavity. The anode lead is brought out through a quarter-wave choke and mica button condenser.

To prevent overheating of the anode of the tube, air must be blown through the cavity. This is done by connecting a low pressure air hose to the air inlet shown in Fig. 1. Excessive air flow must be avoided, as it will cause erratic vibrations of the tuning plunger.

3. Adjustment and Operation

The degree of feedback is adjustable by changing the number and relative positions of the feedback screws which connect the cavity to the grid ring of the tube. There are 16 possible screw positions, but only about 5 or 6 are needed to obtain optimum feedback. Reducing the number of screws increases the amount of feedback.

Care should be taken that the spring which contacts the anode for dc connection is not of such a length to have resonances within the band. When such resonances exist, "holes" or other irregularities will be found in the output spectrum. This spring can act as a helical line, and when it

is too long, resonances will occur which can absorb power and otherwise affect the cavity impedance.

When properly adjusted and sweeping, the output is continuous and the frequency varies approximately sinusoidally back and forth over the band of interest. The width of the sweeping band depends upon the ac current in the voice coil of the speaker drive, and the center frequency depends upon the mean position of the tuning plunger. The latter can be adjusted mechanically by loosening the clamping screw and raising or lowering the sweeping mechanism by hand. It is also possible to make small adjustments of the center frequency electrically by adding a dc component to the voice coil driving current.

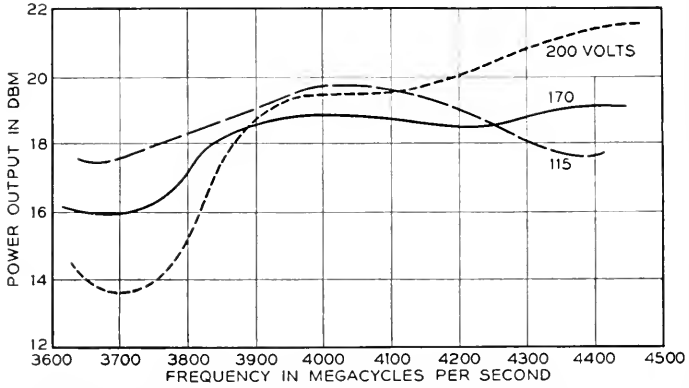

Fig. 2.—Power output curves at 115, 170, and 200 Volts on the anode, for a mean anode current of 25 ma.

Typical curves of power output vs. frequency, taken at different anode voltages, are shown in Fig. 2. The flattest curve requires a voltage considerably lower than the tube rating. The feedback phase is not optimum for best power output, a larger phase shift being desirable in this oscillator. The lowered voltage helps in this regard, increasing the electron transit time in the tube and thereby increasing the phase shift. Efficiency was sacrificed in this design to increase the tuning band. A longer feedback path would increase the power output, but would tend to narrow the band over which oscillation could be obtained by a single tuning adjustment.

The anode power supply should be variable between 100 and 250 volts, but need not be regulated because this voltage is not critical. A rheostat is used for cathode self-bias. The cathode heater and the sweeping mechanism are supplied from a single 6.3 volt filament transformer, with a potentiometer control to vary the sweep range.

A crystal detector and an oscilloscope are used to view the output. It is convenient to use a sinusoidal horizontal sweep on the oscilloscope, driven from the same 6.3 volt transformer as the mechanical sweeping mechanism. In this case, a phase shifter is needed to synchronize the oscilloscope sweep with the motion of the tuning plunger, because there is an appreciable mechanical phase shift in the loudspeaker mechanism.

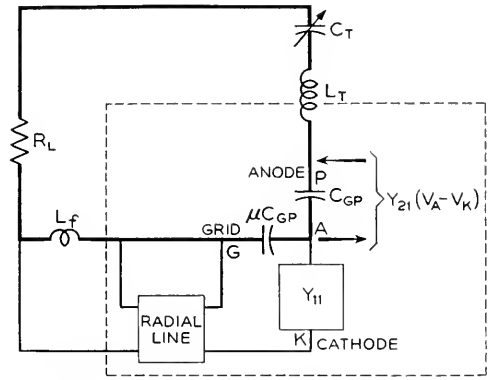

Fig. 3.—Simplified equivalent circuit of the oscillator.

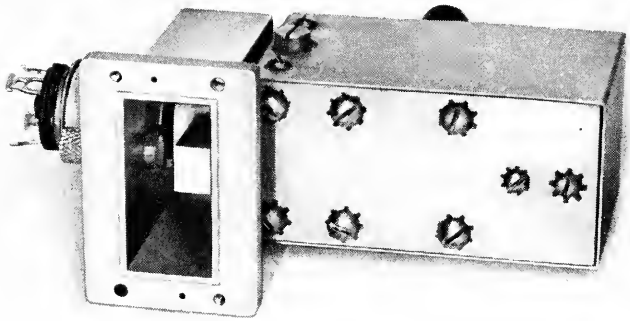

Fig. 4.—The complete oscillator, showing the output coupling window and the ridged waveguide coupling transformer.

When properly phased, the output spectrum will be displayed across the oscilloscope screen with the minimum and maximum frequencies at opposite ends of the trace. In addition, a vibrating relay (such as the Western Electric 275 B Mercury Relay) is used to short out the input to the oscilloscope during half of each cycle. This converts the return trace into a zero-signal reference line, so that the complete picture is a closed loop with a flat bottom. The separation of the active from the reference trace

is a direct indication of signal strength, displayed as a function of frequency.

The results reported here were obtained using the BTL 1553 tube, which is a laboratory model. Samples of the production model, Western Electric 416A, have also been used in this oscillator with quite similar results. To adapt the oscillator for the 416A, the grid ring should be threaded on the inside to fit the threads on the grid disk of that tube.

4. An Equivalent Circuit

The field configuration in the cavity of the oscillator is quite complex, and cannot be readily described in any quantitative fashion. The formulation of an equivalent circuit would require many approximations and

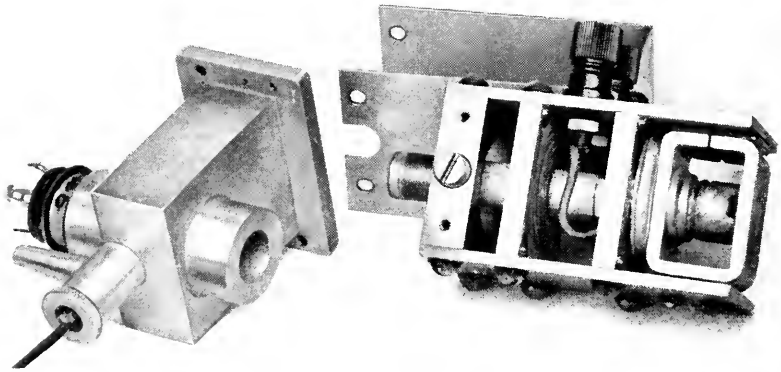

Fig. 5.—The complete oscillator, showing the sweeping mechanism partially dismantled.

judicious guesses if values are to be specified for the various circuit parameters. The circuit of Fig. 3 is believed to be equivalent in a qualitative sense.

A portion of the circuit is within the tube itself. This is the region enclosed by the dotted line in Fig. 3. The T of elements which include Y_{11}, C_{gp}, μC_{gp} and the injected currents, is the equivalent circuit of Llewellyn and Peterson[1] for the active region of a triode. Experimentally determined values for these quantities are reported by Robertson[2]. Y_{11} is the admittance of an equivalent diode between the grid and cathode, and the injected currents indicated by the arrows are the electronic transfer currents associated with the grid voltage and the transadmittance. At high

[1] Llewellyn and Peterson, "Vacuum Tube Networks," *I.R.E. Proc.*, Vol. 32, pp. 144–166 (March 1944).

[2] S. D. Robertson, "Electronic Admittances of Parallel-Plane Electron Tubes at 4000 Megacycles," *B. S. T. J.*, Vol. 28, p. 619, Oct. 1949.

frequencies, both the admittance Y_{11} and the transadmittance Y_{21}, are complex quantities which vary with frequency as shown by Llewellyn and Peterson. The 4-pole box shown represents the passive radial line between the glass seal at the edge of the tube and the cathode-grid gap. This line is heavily loaded with dielectrics and is believed to be electrically about a quarter wavelength long at 4000 Mc. The inductance connected to the anode is that of the anode pin itself and the coaxial center-conductor attached to it. A series resistor R_L is added to include the effects of cavity losses and loading by the output coupling window. C_t is the tuning capacitor which varies with tuning plunger position. The inductance L_f is the feedback reactance introduced by the screws connected to the grid disk.

5. Acknowledgments

I wish to acknowledge the assistance of Messrs. J. A. Morton, R. M. Ryder, and the late A. E. Bowen for many helpful suggestions in the design of this oscillator.

Theory of the Flow of Electrons and Holes in Germanium and Other Semiconductors

By W. VAN ROOSBROECK

(*Manuscript Received Mar. 30, 1950*)

A theoretical analysis of the flow of added current carriers in homogeneous semiconductors is given. The simplifying assumption is made at the outset that trapping effects may be neglected, and the subsequent treatment is intended particularly for application to germanium. In a general formulation, differential equations and boundary-condition relationships in suitable reduced variables and parameters are derived from fundamental equations which take into account the phenomena of drift, diffusion, and recombination. This formulation is specialized so as to apply to the steady state of constant total current in a single cartesian distance coordinate, and properties of solutions which give the electrostatic field and the concentrations and flow densities of the added carriers are discussed. The ratio of hole to electron concentration at thermal equilibrium occurs as parameter. General solutions are given analytically in closed form for the intrinsic semiconductor, for which the ratio is unity, and for some limiting cases as well. Families of numerically obtained solutions dependent on a parameter proportional to total current are given for *n*-type germanium for the ratio equal to zero. The solutions are utilized in a consideration of simple boundary-value problems concerning a single plane source in an infinite filament.

Table of Contents

1. Introduction

IN A semiconductor there are current carriers of two types: electrons in the conduction band, and positive holes in the filled valence band; and the increase of their concentrations in the volume of the semicon-

ductor over the concentrations which obtain at thermal equilibrium is fundamental to a number of related phenomena, of which transistor action is a familiar instance. In an n-type semiconductor, for example, in which the carriers are predominantly electrons, the carrier concentrations are increased by the introduction of holes which, through a process of space-charge neutralization, produce additional electrons in the same numbers and concentrations. The bulk conductivity of the semiconductor is thereby so increased that power gain is obtainable.[1] Holes can be introduced by the local application of heat, or by irradiation with light, X-rays, or high-velocity electrons—in fact, by any agency which transfers electrons from the highest filled band to the conduction band. They can be introduced also through an emitter, which may be a positively biased point contact or a positively biased $p - n$ junction[3], as exemplified in the transistor. In this case the emitter introduces holes, which flow into the volume of the semiconductor[4], by the removal of electrons from the filled band.[2, 5] Entirely analogous considerations apply to the introduction of electrons into a p-type semiconductor.[6]

In their flow in a semiconductor, added electrons and holes are subject to drift under electrostatic fields and to diffusion in the presence of concentration gradients as a consequence of their random thermal motions. They are subject also to recombination, which results in concentration gradients in source-free regions even for the steady state in one dimension, or which augments those which may otherwise be associated with the time-dependence of the flow, or with its geometry in the steady state. From fundamental equations which take into account these phenomena of drift, diffusion, and recombination, for the existence of each of which there is experimental evidence[1], general differential equations and boundary-condition relationships in suitable reduced or dimensionless variables and parameters may be derived, and solutions which give the concentrations and flow densities of added carriers obtained for various cases of physical interest.

This paper presents results of a theoretical analysis, along these lines, of the flow of electrons and holes in semi-conductors. The treatment is intended particularly for application to germanium. An initial formulation,

[1] W. Shockley, G. L. Pearson and J. R. Haynes, *B. S. T. J.* 28, (3), 344–366 (1949).
[2] J. Bardeen and W. H. Brattain, *Phys. Rev.* 74 (2), 230–231 (1948); W. H. Brattain and J. Bardeen, *Phys. Rev.* 74 (2) 231–232 (1948).
[3] W. Shockley, G. L. Pearson and M. Sparks, *Phys. Rev.* 76 (1), 180 (1949); W. Shockley, *B. S. T. J.* 28 (3), 435–489 (1949).
[4] E. J. Ryder and W. Shockley, *Phys. Rev.* 75 (2), 310 (1949); J. N. Shive, *Phys. Rev.* 75 (4), 689–690 (1949); J. R. Haynes and W. Shockley, *Phys. Rev.* 75 (4), 691 (1949).
[5] J. Bardeen and W. H. Brattain, *Phys. Rev.* 75 (8), 1208–1225 (1949); *B. S. T. J.* 28 (2), 239–277 (1949).
[6] W. G. Pfann and J. H. Scaff, *Phys. Rev.* 76 (3), 459 (1949); R. Bray, *Phys. Rev.* 76 (3), 458 (1949).

which retains, wherever convenient, such generality as is instructive per se or of manifest utility, is specialized so as to apply to the steady state of constant current in a single cartesian distance coordinate. For the intrinsic semiconductor, general analytical solutions are obtainable in closed form, and such solutions are given, as well as general solutions obtained numerically for n-type germanium in which the hole concentration at thermal equilibrium may be neglected compared to the electron concentration. Solutions for these cases are given explicitly for each of two recombination laws: recombination according to a mass-action law, and recombination such that the mean lifetime of the added carriers is constant. Methods are described for the fitting of boundary conditions, and the following relatively simple boundary-value problems are considered: a source at the end of a semi-infinite semi-conductor filament; and a single source in a doubly-infinite filament.

To indicate the presumed scope and application of the results obtained, it may suffice to outline briefly the principal assumptions on which they are based and the approximations employed: The assumption is made at the outset that trapping effects may be neglected, which provides the important simplification that the recombination rates of holes and electrons are equal at all times. One justification for this is the circumstance that the fairly high hole mobilities found by G. L. Pearson from Hall-effect and conductivity measurements[7] are no larger than those found by J. R. Haynes from transit times under pulse conditions[1]. With hole trapping, holes injected in a pulse would initially fill traps; and if there were subsequent relatively slow release of the holes from the traps, an apparent reduction of mobility would be manifest. It is further assumed that substantially all donor and acceptor impurities are ionized. With the assumption that the semi-conductor is homogeneous in its bulk, and free from grain boundaries[8] or rectifying barriers, the assumption of the electrical neutrality of the semiconductor, or of the neglect of space charge, is in general an excellent approximation: Small departures from electrical neutrality in the volume would vanish rapidly, with time constant equal to that for the dielectric relaxation of charge, which for germanium equals $1.5 \cdot 10^{-12}$ sec per ohm cm of resistivity[9] and is in general small compared with the mean lifetime of added carriers. A uniform local departure from electrical neutrality in germanium of only one per cent in relative concentration would produce appreciable changes in field in a

[7] G. L. Pearson, *Phys. Rev.* 76 (1), 179–180 (1949).

[8] G. L. Pearson, *Phys. Rev.* 76 (3), 459 (1949); W. E. Taylor and H. Y. Fan, paper OA5, and N. H. Odell and H. Y. Fan, paper OA2 of the 1950 Annual Meeting of the American Physical Society, February 3, 1950.

[9] A value of 16.6 for the dielectric constant of germanium is obtained from optical data of H. B. Briggs: *Phys. Rev.* 77 (2), 287 (1950).

mean free path for the carriers, equal to $1.1 \cdot 10^{-5}$ cm at room temperature, which would even preclude the applicability of the fundamental equations employed. In qualitative terms, the conductivity of the semiconductor is sufficiently large that the currents which commonly occur are produced by moderate fields whose maximum gradients are relatively small. Space charge may persist in the steady state, but then only in surface regions whose thickness[10] in germanium is generally less than about 10^{-4} cm and whose effects may be dealt with through suitable boundary conditions.

The steady-state solutions, in their qualitative aspects, are illustrative of the phenomena taken into consideration. In an extrinsic semiconductor, if the concentrations of added carriers are not too large, the solutions for moderate and large fields are in general approximately ohmic in their local behavior. The effect of diffusion is then comparatively small, and the added carriers largely drift under a field which varies with distance through the increased conductivity which these recombining carriers themselves produce. Diffusion effects are incident in addition to this behavior, and become pronounced for large concentrations or small applied fields. For example, solutions which specify the concentrations of added holes as functions of distance, for different total currents or applied fields in a source-free region, all approach a common solution for large hole concentrations, regardless of applied field; those for the hole current and the electrostatic field behave similarly. This behavior results from diffusion in conjunction with the increase in conductivity. Another example is that of the solutions for zero total current: As the result of diffusion in conjunction with recombination, a flow of added holes can occur along a semi-conductor filament with no flow of current. It is, of course, accompanied by an equal electron flow, so that the hole and electron currents cancel, and occurs in any open-circuited semi-conductor filament which adjoins a region in which added holes flow. It can also be realized by suitable irradiation of an end of a filament, with no applied field. A closely related effect is illustrated in the flow of holes injected through a point-contact emitter into a semi-conductor filament along which a sweeping field is applied: Some of the holes will flow against the field, an appreciable proportion, unless the current in the filament is sufficiently large. As a further example, if the mobilities of holes and electrons were equal, the electrostatic field would be given by Ohm's law as the total current

[10] The (largest) distance over which the increment in electrostatic potential exceeds kT/e may be expressed in units of the length $L_d \equiv (kT\epsilon/8\pi n_i e^2)^{\frac{1}{2}}$, where n_i is the thermal-equilibrium concentration of electrons (or holes) in the intrinsic semiconductor; see the paper of reference 3, also W. Schottky and E. Spenke, *Wiss. Veröff. Siemens-Werken* 18 (3), 1–67 (1939). This distance increases with resistivity, never exceeding the value 1.4 L_d for the intrinsic semiconductor. In high back voltage n-type germanium, it exceeds about 0.5 L_d, and L_d for germanium is about $7.4 \cdot 10^{-5}$ cm at room temperature.

divided by the local increased conductivity. With electrons more mobile than holes, this ohmic field is modified by a contribution which is directed away from a hole source and proportional to the magnitude of the concentration gradient divided by the local conductivity. This contribution gives a non-vanishing electrostatic field for zero total current.

The intrinsic semiconductor has, as the result of a conductivity which is everywhere proportional to the concentration of carriers of either type, the property that the flow in it is as if the added carriers were actuated entirely by diffusion, with only the carriers normally present drifting under a field equal to the unmodulated applied field. The extrinsic semiconductor becomes in effect intrinsic if the concentrations of carriers are sufficiently increased, by whatever means, the ohmic contribution to the current density of either electrons or holes then becoming proportional to the total current density and, in this case, negligible compared with the contribution due to diffusion. It may, for example, be expected that the transport velocity of added carriers in an extrinsic semiconductor can be increased by an increase in the applied field only if the consequent joule heating does not unduly modify the semiconductor in the intrinsic direction.

General solutions for the steady state in one dimension are obtainable analytically in closed form for a number of important special cases. Aside from that for which diffusion is neglected, they include the general cases for no recombination, for the intrinsic semiconductor, and for zero total current, and the limiting cases of small and of large concentrations of added carriers. W. Shockley has made use of small-concentration theory in an analysis of $p - n$ junctions[3]. J. Bardeen and W. H. Brattain have given solutions for the steady-state hole flow in three dimensions, neglecting recombination, in the neighborhood of a point-contact emitter.[5, 11] Transient solutions are obtainable analytically for the intrinsic semiconductor for constant mean lifetime, and for the extrinsic semiconductor if the concentrations of added carriers are sufficiently small that the change in conductivity is negligible. For concentrations unrestricted in magnitude, Conyers Herring has described a general method for graphical or numerical construction of transient solutions in one dimension from a first-order partial differential equation appropriate to the case for which diffusion is neglected in the extrinsic semi-conductor, and has given some solutions so obtained, with estimates of the effect of diffusion. Reference might be made to his paper[12] also for discussion of various physical con-

[3] loc. cit.
[5] loc. cit.
[11] See the paper of J. Bardeen in this issue.
[12] Conyers Herring, *B. S. T. J.* 28 (3), 401–427 (1949).

siderations and of certain interesting transient effects. Steady-state alternating-current theory for relatively small total hole concentrations in the n-type semiconductor has been used to describe the action of the filamentary transistor[13] for which diffusion may in general be neglected.[1]

The steady-state solutions in one dimension apply to single-crystal semiconductor filaments, and for critical comparisons between theory and experiment, the ideal one-dimensional geometry should be simulated as closely as possible. Experimental estimates of hole concentrations and flows are frequently obtained from measurements of potentials and conductances of point contacts along a filament[1]. These estimates require a knowledge of the dependence of the current-voltage characteristics of point contacts on hole concentration. Theory for this dependence has been presented by J. Bardeen[11], and the determination of hole concentrations by means of the solutions here given should provide an essential adjunct to this point contact theory for its comparison with experiment.

2. General formulation

2.1 *Outline*

The formulation of the general problem is initiated by writing the fundamental equations for the time-dependent flow of holes and electrons in a source-free region of a homogeneous semiconductor under the assumption that there is no trapping. Conditions for their validity are discussed. Neglecting changes in the concentrations of ionized donors and acceptors, the fundamental equations are expressed in reduced or dimensionless form by suitable transformations of the dependent and independent variables. They are simplified so that the general problem is formulated by means of second-order partial differential equations in two dependent variables, one for concentration and the other for electrostatic potential; corresponding equations are derived for the intrinsic semiconductor. Various properties of the equations are adduced. For the flow in one dimension, a differential equation in the hole concentration is given for the n-type semiconductor, accompanied by expressions for the electrostatic field and hole flow density, as well as by some boundary-condition relationships involving specification of the latter. The equations for this case are found to depend on three parameters: the ratio of electron to hole mobility; a reduced concentration of holes at thermal equilibrium; and a parameter which fixes the total current density.

The recombination of holes and electrons is specified by means of a

[1] loc. cit.
[11] loc. cit.
[13] W. Shockley, G. L. Pearson, M. Sparks, and W. H. Brattain, *Phys. Rev.* 76 (3), 459 (1949).

suitable function of the concentration of the added carrier, whose form is specified for two recombination laws: recombination according to a mass-action law, and recombination characterized by constant mean lifetime. It is shown that essentially the same reduced equations apply to the case for which recombination is neglected.

Second-order differential equations in the hole concentration for the n-type semiconductor with the thermal-equilibrium value of the hole concentration assumed negligible compared to the electron concentration, and for the intrinsic semiconductor, are then written for the steady state of constant current in one dimension. These are converted into first-order equations which have, as dependent variable a reduced concentration gradient G, and as independent variable a reduced concentration of added holes, ΔP. Boundary conditions are expressed as relationships between these variables. Properties of the general solutions and of the boundary conditions are accordingly examined in the $(\Delta P, G)$-plane. It is found that there are two intersecting solutions through the $(\Delta P, G)$-origin, which is a saddle-point of the differential equation, and that these are the solutions for field directed respectively towards and away from sources in semi-infinite regions which have sources only to one side. They are called field-opposing and field-aiding solutions, and possess two degrees of freedom. Solutions which do not intersect at the origin are asymptotic to these, possess three degrees of freedom, and are called solutions of the composite type. This is the general type, and applies to a finite region in distance at both ends of which boundary conditions are specified. The region may, for example, be one between a source and either another source, a sink, a non-rectifying electrode, or a surface upon which recombination takes place. While the analysis of composite cases is straightforward, the present treatment is confined to the simpler cases of field opposing and field aiding, the latter being the one most generally applicable to experiments in hole injection. Also, where the differential equations involved are linear, solutions for composite cases can be written as linear combinations of field-aiding and field-opposing solutions.

From the properties of the curves in the $(\Delta P, G)$-plane is determined the qualitative behavior of the hole concentration at a hole source at the end of a semi-infinite filament as the total current is indefinitely increased.

2.2 *Fundamental equations for the flow of electrons and holes*

The equations for the flow in three dimensions of electrons and holes in a homogeneous semiconductor contain, as principal dependent variables, the hole and electron concentrations, p and n, the flow densities $\mathbf{J}_p$ and $\mathbf{J}_n$, and the electrostatic field, $\mathbf{E}$, or potential, V. With no trapping,

the equations may be written in a symmetrical form, so that they are applicable to either an n-type, a p-type, or an intrinsic semiconductor, as follows:

$$
(1)\qquad
\begin{cases}
\dfrac{\partial p}{\partial t} = -\left[p/\tau_p - g_0\right] - \operatorname{div} \mathbf{J}_p \\[2ex]
\dfrac{\partial n}{\partial t} = -\left[n/\tau_n - g_0\right] - \operatorname{div} \mathbf{J}_n \\[2ex]
\mathbf{J}_p \equiv \dfrac{1}{e}\mathbf{I}_p = \mu_p\left[p\mathbf{E} - \dfrac{kT}{e}\operatorname{grad} p\right] = -\mu_p p \operatorname{grad}\left[V + \dfrac{kT}{e}\log p\right] \\[2ex]
\mathbf{J}_n \equiv -\dfrac{1}{e}\mathbf{I}_n = \mu_n\left[-n\mathbf{E} - \dfrac{kT}{e}\operatorname{grad} n\right] \\[2ex]
\qquad\qquad = -\mu_n n \operatorname{grad}\left[-V + \dfrac{kT}{e}\log n\right] \\[2ex]
\operatorname{div} \mathbf{E} = \dfrac{4\pi e}{\epsilon}\left[(p - p_0) - (n - n_0) + (D^+ - D_0^+) - (A^- - A_0^-)\right] \\[2ex]
\mathbf{E} = -\operatorname{grad} V.
\end{cases}
$$

In the first two equations, which are the continuity equations for holes and electrons written for a region free from external sources, g_0 is a constant which represents the thermal rate of generation of hole-electron pairs per unit volume; for cases in which hole-electron pairs are produced also by penetrating radiation, appropriate source terms in the form of identical functions of the space and time coordinates can be included on the right in the respective equations. The mean lifetimes of holes and electrons, τ_p and τ_n, are in general considered to be concentration-dependent and, since trapping is neglected, the quantities p/τ_p and n/τ_n are equal, being the rate at which holes and electrons recombine. Evaluated for the normal semiconductor, or the semiconductor at thermal equilibrium with no injected carriers, they equal g_0.

The equations for $\mathbf{J}_p$ and $\mathbf{J}_n$, which are vectors whose magnitudes equal, respectively, the numbers of holes and of electrons which traverse unit area in unit time, are diffusion equations of M. von Smoluchowski, written for hole flow and for electron flow[14]. Of the type frequently employed, after C. Wagner, in theories of rectification, each expresses the dependence of the flow density on the electrostatic field and on the concentration gradient, the diffusion constant for holes or electrons having been expressed in terms of the mobility, μ_p or μ_n, in accordance with the

[14] S. Chandrasekhar, *Rev. Mod. Phys.* 15, 1–89 (1943).

well-known relationship of A. Einstein[15]. In them, e denotes the magnitude of the electronic charge; T is temperature in degrees absolute; and k is Boltzmann's constant. With transport velocity defined as flow density divided by concentration, the product of the mobility and the quantity in square brackets in the expression for $\mathbf{J}_p$ or $\mathbf{J}_n$ on the extreme right gives the corresponding velocity potential, which is thus proportional to the sum of an electrostatic potential and a diffusion potential.

The next to last equation is Poisson's equation, which relates the divergence of the field to the net electrostatic charge. Here ϵ is the dielectric constant; p_0 and n_0 are the concentrations of holes and electrons at thermal equilibrium, in the normal semiconductor. The concentrations of ionized donor and acceptor impurities at thermal equilibrium are represented by D_0^+ and A_0^-, while D^+ and A^- are dependent variables which denote the respective concentrations in general of ionized donors and acceptors in the semiconductor with added carriers. As shown in the Appendix, variations in D^+ and A^- may be neglected if the impurity centers are substantially all ionized in the normal semiconductor, despite the effect large concentrations of added carriers may have on the equilibria[16].

The expression of the electrostatic field as the gradient of a potential according to the last equation is consistent with the circumstance that the effects of magnetic fields, with none applied, are in general quite negligible.

Subtracting the first continuity equation from the second, it is found that

$$\text{div } (\mathbf{J}_p - \mathbf{J}_n) = -\frac{\partial}{\partial t}(p - n), \tag{2}$$

since, with no trapping, p/τ_p equals n/τ_n. Neglecting changes in the concentrations of ionized donors and acceptors, this equation and Poisson's equation give

$$\mathbf{J}_p - \mathbf{J}_n = \mathbf{J} - \frac{\epsilon}{4\pi c}\frac{\partial \mathbf{E}}{\partial t}; \qquad \mathbf{I}_p + \mathbf{I}_n = \mathbf{I} - \frac{\epsilon}{4\pi}\frac{\partial \mathbf{E}}{\partial t}, \tag{3}$$

where $\mathbf{J}$ and $\mathbf{I}$ are solenoidal vector point functions, in general time-dependent. The latter is the total current density, and the term which follows it in (3) gives the displacement current density.

[15] A. Einstein, *Annalen der Physik* 17, 549–560 (1905); Müller-Pouillet, *Lehrbuch der Physik*, Braunschweig, 1933, IV (3), 316–319.

[16] It has been found from measurements of the temperature dependence of the conductivity and Hall coefficient that the energy of thermal ionization of the donors in n-type germanium of relatively high purity is only about $10^{-2}eT$, whence most of the donors are ionized at room temperature: G. L. Pearson and W. Shockley, *Phys. Rev.* 71 (2), 142 (1947).

It may be well to point out that the validity of the diffusion equations depends on two assumptions, which, while hardly restrictive in general for homogeneous semiconductors, indicate the nature of the generalizations which might otherwise be necessary[14]. The first assumption is that there are no appreciable time changes in the dependent variables in the relaxation time for the conductivity, or the time of the elementary fluctuations. This is tantamount to the requirement that the carriers undergo many collisions in the time intervals of interest. The second assumption is that the changes in the carriers' electrostatic potential energy over distances equal to the mean free path are small compared with the average thermal energy. In accordance with this assumption, very large fields in the electrically neutral semiconductor for which the carriers are not substantially in thermal equilibrium with the lattice are ruled out. The neglect of space charge then in general validates the two assumptions, if the resistivity is not too small, since the neglect of changes in the dependent variables which occur in the dielectric relaxation time obviates their change in the relaxation time for conductivity; and the neglect in the steady state of appreciable variations in electrostatic potential, and thus in the other dependent variables, in the distance[10] L_d, obviates their variation in a mean free path. The dielectric relaxation time for germanium, $1.5 \cdot 10^{-12}$ sec per ohm cm of resistivity, in high back voltage material exceeds the relaxation time for conductivity, which is about $1.0 \cdot 10^{-12}$ sec; and in semi-conductors in which the mobilities and the conductivity are smaller than the comparatively large values for germanium, the dielectric relaxation time may be appreciably larger than the relaxation time for conductivity. Similarly, L_d for germanium is about 7 times the mean free path, and this ratio, which is essentially inversely proportional to the square root of the product of mobility and intrinsic conductivity, may be appreciably larger for other semiconductors.

If, on the other hand, it should be desired to consider space-charge effects in germanium, the diffusion equations may be of rather marginal applicability, and the use of their appropriate generalization indicated, since with L_d equal to 7 mean free paths, appreciable space-charge variation of potential, corresponding to a field which is not small compared with the free-path thermal-energy equivalent of about 3500 volt cm^{-1}, may occur in at least one of the free paths. For example, diode theory, rather than diffusion theory, provides the better approximation for the characteristics of germanium point-contact rectifiers, and is particularly applicable to those from low resistivity material for which the potential variation is largely confined to one mean free path or less[17].

[10] loc. cit.
[14] loc. cit.
[17] H. C. Torrey and C. A. Whitmer, "Crystal Rectifiers," New York, 1948, Sec. 4.3.

Neglecting space charge, Poisson's equation becomes simply the condition of electrical neutrality:

$$(p - p_0) - (n - n_0) = 0, \tag{4}$$

assuming substantially complete ionization of donors and acceptors. Similarly, equations (3) become

$$\mathbf{J}_p - \mathbf{J}_n = \mathbf{J}; \qquad \mathbf{I}_p + \mathbf{I}_n = \mathbf{I}. \tag{5}$$

With electrical neutrality, the two continuity equations merge into one: Since derivatives of p equal the corresponding ones of n,

$$\begin{aligned} \operatorname{div} \mathbf{J}_p &= -[p/\tau_p - g_0] - \frac{\partial p}{\partial t} \\ &= \operatorname{div} \mathbf{J}_n = -[n/\tau_n - g_0] - \frac{\partial n}{\partial t}. \end{aligned} \tag{6}$$

The neutrality condition in conjunction with the two equations obtained by substituting for $\mathbf{J}_p$ and $\mathbf{J}_n$ from the diffusion equations in (6) thus provide three equations for the determination of p, n, and $\mathbf{E}$ or V.

It is instructive to rewrite equations (6) in accordance with

$$\begin{cases} \operatorname{div} \mathbf{J}_p = \mathbf{s} \cdot \operatorname{grad} p \\ = \operatorname{div} \mathbf{J}_n = \mathbf{s} \cdot \operatorname{grad} n, \\ \mathbf{s} \equiv \left[\dfrac{\partial \mathbf{J}_p \cdot \mathbf{i}}{\partial x} \Big/ \dfrac{\partial p}{\partial x}\right] \mathbf{i} + \left[\dfrac{\partial \mathbf{J}_p \cdot \mathbf{j}}{\partial y} \Big/ \dfrac{\partial p}{\partial y}\right] \mathbf{j} + \left[\dfrac{\partial \mathbf{J}_p \cdot \mathbf{k}}{\partial z} \Big/ \dfrac{\partial p}{\partial z}\right] \mathbf{k}, \end{cases} \tag{7}$$

where $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors in the directions of the respective axes. The velocity $\mathbf{s}$, which is given as well by the expression for electrons analogous to that written for holes, may be defined alternatively as follows: Suppose, for definiteness, that the second-order system of equations (4) and (6) have been solved, so that the concentrations and flow densities are known in terms of the cartesian coordinates x, y, and z, and the time t. The x-component of $\mathbf{s}$ is then the partial derivative with respect to p of the x-component of $\mathbf{J}_p$ in which x has been replaced by the proper function of p, y, z, and t, and similarly for the other components. Thus, with $\mathbf{s}$ a known function, p or n may be considered to satisfy the first-order partial differential equation obtained by substituting from (7) in (6), from which it is evident that $\mathbf{s}$ is the velocity with which concentration transients are propagated[18]. This velocity, which is here called the differential

[18] The identification of $\mathbf{s}$ as this propagation velocity follows the example of C. Herring, in whose method for solving the transient constant-current problem in one dimension the velocity depends in a known manner on concentration only, through the neglect of diffusion, so that the general solution of the differential equation in which thus neither independent variable x nor t occurs explicitly may be obtained; cf. reference 12, pp. 412 ff.

transport velocity and loosely referred to as the transport velocity of added carriers, of course differs in general from the transport velocity proper, defined as the ratio of flow density to concentration; its general definition, which is applicable to the steady state, has been introduced to facilitate later interpretations.

2.3 *Reduction of the fundamental equations to dimensionless form*

2.31 *The general case*

In order to obtain solutions in forms which exhibit such generality as they may possess, the fundamental equations are to advantage written in terms of dimensionless dependent and independent variables which are the original variables measured in suitable units. Through formal consideration of the equations (1), in conjunction with (3) or with (4) and (6), these units can be so chosen that the system of reduced equations will exhibit independent parameters on which it may be considered to depend. The best choice of suitable units is by no means unique; those choices which have been made are natural ones, in that they have been found to result in greater formal simplicity and ease of interpretation in the theory than others which may be equally valid in principle.

The choice for an n-type semiconductor consists in definitions of dimensionless variables and parameters as follows:

$$
\begin{aligned}
&X \equiv x/L_p,\ Y \equiv y/L_p,\ Z \equiv z/L_p;\ L_p \equiv \left[\frac{kT\mu\tau}{e}\right]^{\frac{1}{2}} = [D_p\tau]^{\frac{1}{2}} \\
&U \equiv t/\tau \\
&P \equiv p/(n_0 - p_0);\ P_0 \equiv p_0/(n_0 - p_0) = g_0\tau/(n_0 - p_0) \\
&N \equiv n/(n_0 - p_0);\ N_0 \equiv n_0/(n_0 - p_0) \\
&\mathbf{C} \equiv \mathbf{I}/I_0 = \mathbf{E}_a/E_0 = \mu\mathbf{E}_a/[D_p/\tau]^{\frac{1}{2}};\ I_0 \equiv \sigma_0 E_0;\ E_0 \equiv kT/eL_p \\
&\mathbf{C}_p \equiv \mathbf{I}_p/I_0 \\
&\mathbf{C}_n \equiv -\mathbf{I}_n/I_0 \\
&\mathbf{F} \equiv \mathbf{E}/E_0 \\
&W \equiv V/E_0L_p = eV/kT \\
&Q \equiv \tau/\tau_p\,.
\end{aligned}
\tag{8}
$$

The rectangular cartesian space coordinates are x, y, and z. The quantity τ is the mean lifetime of holes for concentrations of added holes small compared with the thermal-equilibrium electron concentration, n_0; and σ_0 is the conductivity of the normal semiconductor. The hole mobility,

originally μ_p, is denoted by μ for simplicity. If b is the ratio of electron to hole mobility, σ_0 is given in general by

$$(9) \quad \sigma_0 = \mu e(bn_0 + p_0) = M_0\, b\mu e(n_0 - p_0), \qquad M_0 \equiv 1 + \frac{b+1}{b} P_0,$$

the symbol M_0 being introduced for brevity. If $p_0 << n_0$, M_0 is unity and σ_0 equals $b\mu e n_0$.

The independent dimensionless distance variables are X, Y and Z, where the distance unit, L_p, is a diffusion length for a hole for the mean lifetime, τ, the diffusion constant for holes being D_p. This mean lifetime is the unit for the independent dimensionless time variable, U. The hole and electron concentrations are measured in units of the excess in concentration of electrons over holes[19], $n_0 - p_0$, the reduced variables being P and N, respectively. The reduced total current $\mathbf{C}$ is total current density measured in units of the current density I_0 which flows in the semiconductor with no added carriers under the characteristic field E_0, which is a field such that a carrier would expend the energy kT in drifting with it through the distance L_p. A more illuminating alternative description is that $\mathbf{C}$ is the ratio of the average drift velocity of holes under the applied or asymptotic field, $\mathbf{E}_a$, to the hole diffusion velocity $(D_p/\tau)^{\frac{1}{2}}$. The field $\mathbf{E}_a$ is that which produces the current density $\mathbf{I}$ in the semiconductor with no added carriers. The corresponding reduced hole and electron flow densities are $\mathbf{C}_p$ and $\mathbf{C}_n$. The electrostatic field measured in units of E_0 is denoted by $\mathbf{F}$, and W is the corresponding reduced electrostatic potential. The lifetime ratio Q is a function of P which characterizes the recombination process. While it appears from experiment that the recombination rate for holes depends on both physical and chemical properties of the semiconductor, in a particular semiconductor at given temperature it may be considered to depend on hole concentration alone.

Representative values for germanium of units in terms of which the dimensionless quantities are defined are as follows: The mean lifetime τ may be of the order of 10^{-5} sec. With a mobility for holes[7] of 1700 cm^2 volt^{-1} sec^{-1} in germanium single crystals at 300 deg abs, the length L_p is then about $2 \cdot 10^{-2}$ cm; the characteristic field, E_0, 1.2 volt cm^{-1}; and the current density I_0, 0.12 ampere cm^{-2} for a resistivity of 10 ohm cm.

With these definitions[20], the fundamental equations for a region free from external sources, neglecting changes in the concentrations of ionized

[7] loc. cit.

[19] The excess in concentration of electrons over holes is of course equal to that of ionized donors over ionized acceptors.

[20] The definitions given appear best if there is a region in which $P - P_0$ is small, with $P_0 \neq 0$. Modified definitions of the reduced flow densities, in which the conductivity σ_0 is replaced by the conductivity $b\mu e(n_0 - p_0)$ due to the excess electrons alone, result in equations obtainable formally by setting M_0 equal to unity.

donors and acceptors and neglecting space charge[21], are given in reduced form as follows:

$$(10)\quad \begin{cases} \dfrac{\partial P}{\partial U} = -[bM_0 \operatorname{div} \mathbf{C}_p + PQ - P_0] \\ \dfrac{\partial N}{\partial U} = -[bM_0 \operatorname{div} \mathbf{C}_n + PQ - P_0] \\ \mathbf{C}_p = \dfrac{1}{bM_0}[\mathbf{F}P - \operatorname{grad} P] = -\dfrac{1}{bM_0} P \operatorname{grad}[W + \log P] \\ \mathbf{C}_n = \dfrac{1}{M_0}[-\mathbf{F}N - \operatorname{grad} N] = -\dfrac{1}{M_0} N \operatorname{grad}[-W + \log N] \\ (P - P_0) - (N - N_0) = P - N + 1 = 0 \\ \mathbf{F} = -\operatorname{grad} W, \end{cases}$$

and the reduced form of equations (5) is

$$(11)\qquad \mathbf{C}_p - \mathbf{C}_n = \mathbf{C}.$$

These reduced equations may be simplified and two differential equations in the dependent variables P and W written as follows:

$$(12)\quad \begin{cases} -b M_0 \operatorname{div} \mathbf{C}_p = \operatorname{div} P \operatorname{grad}[W + \log P] = [PQ - P_0] + \dfrac{\partial P}{\partial U} \\ \operatorname{div} \mathbf{C} = 0, \qquad \mathbf{C} = -\Sigma \operatorname{grad}\left[W - \dfrac{b-1}{b+1} \log \Sigma\right], \end{cases}$$

where Σ is the conductivity σ in reduced form:

$$(13)\qquad \Sigma \equiv \frac{\sigma}{\sigma_0} = \frac{bN + P}{bN_0 + P_0} = \frac{1}{M_0}\left[1 + \frac{b+1}{b} P\right].$$

An alternative formulation, due to R. C. Prim, which is obtained by evaluating div $[\mathbf{C}_n \pm b\, \mathbf{C}_p]$, consists of the two equations,

$$(14)\quad \begin{aligned} \frac{b}{b-1} \operatorname{div}(1 + 2P) \operatorname{grad} W &= -\frac{b}{b+1} \operatorname{div} \operatorname{grad}[W - (1 + 2P)] \\ &= [PQ - P_0] + \frac{\partial P}{\partial U}, \end{aligned}$$

[21] It may be desirable to take space charge into account in cases involving high frequencies or high resistivities. Poisson's equation and equations (3) are in reduced form,

$$P - N + 1 = bM_0 \Gamma \operatorname{div} \mathbf{F} \text{ and } \mathbf{C}_p - \mathbf{C}_n = \mathbf{C} - \Gamma \frac{\partial \mathbf{F}}{\partial U}, \text{ where } \Gamma \equiv \epsilon / 4\pi\sigma_0 \tau.$$

The term containing Γ may often be omitted from one of these equations, depending on the nature of the particular case considered.

in which the use of $(1 + 2P)$ as dependent variable may be desirable. This variable is equal to the concentration of carriers of both kinds divided by the excess of electron concentration over hole concentration, which is a constant.

The expression in the equations which specifies the recombination rate may be written more simply. Since the lifetime ratio Q is unity for $P = P_0$,

$$PQ - P_0 = (P - P_0)R, \tag{15}$$

where R, which will be called the recombination function, depends on P and also equals unity for $P = P_0$. The lifetime ratio and the recombination function which, of course, differ in general, both equal unity for the case of constant mean lifetime. Recombination of holes and electrons at a rate proportional to the product of their concentrations, called mass-action recombination, and recombination characterized by a constant mean lifetime for holes are frequently of interest. For a combination of independent mechanisms of both types, it is easily seen that

$$\begin{cases} Q \equiv \tau/\tau_p = 1 + a\,(p - p_0)/n_0 = 1 + a\,(P - P_0)/(1 + P_0), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad a \equiv \tau/\tau_v,\ 0 \leq a \leq 1 \\ R = 1 + ap/n_0 = 1 + aP/(1 + P_0), \end{cases} \tag{16}$$

where τ_v is the mean lifetime for small concentrations associated with mass-action recombination alone, so that $a = 0$ for constant mean lifetime, and $a = 1$ for mass-action recombination. If both recombination mechanisms are operative, that of mass-action recombination will, of course, determine the mean lifetime where the concentration of added carriers is sufficiently large.

Recent experiments have shown that the mean lifetime for holes in n-type germanium can be increased materially, to at least 100 microseconds, by minimizing surface recombination through decreases in surface-to-volume ratios.[1] On the other hand, comparatively short mean lifetimes, of the order of one microsecond, occur in p-type germanium produced, for example, from n-type by nucleon bombardment. It should be possible to determine in various cases which recombination law would provide the better approximation by use of the technique of H. Suhl and W. Shockley of hole injection in the presence of a magnetic field[22] or by the photoelectric technique of F. S. Goucher[23].

[1] loc. cit.

[22] H. Suhl and W. Shockley, *Phys. Rev.* 75 (10), 1617–1618; 76 (1), 180 (1949).

[23] F. S. Goucher, paper I 11 of the Oak Ridge Meeting of the American Physical Society, March 18, 1950; *Phys. Rev.* 78 (6), 816 (1950).

It appears that solutions neglecting recombination furnish useful approximations for some applications. If recombination is neglected, by assuming that the mean lifetime is infinite, the definitions (8) of the dimensionless quantities no longer have meaning, but essentially the same differential equations and corresponding boundary-condition equations can still be used. The reduced equations become essentially homogeneous in τ for τ large, and it suffices to suppress the recombination terms, $PQ - P_0$, retaining formally the definitions of the dimensionless quantities in which now τ, and thus L_p and E_0 or I_0 no longer have physical significance. One of these unitary quantities may be chosen arbitrarily. It might be noted that if Poisson's equation is retained the length unit is advantageously chosen as L_d, which gives a dielectric relaxation time for the time unit.[21]

In one cartesian dimension, with total current a function of time only, H' may be eliminated by means of the equation for C in (12) and, upon substituting for it in any of the three remaining equations in (12) and (14), a differential equation for P results which depends on b, P_0, and C as parameters. Dropping vector notation, this equation is

$$\frac{\partial P}{\partial U} = \frac{\left[(1+2P)\left(1+\frac{b+1}{b}P\right)\right]\frac{\partial^2 P}{\partial X^2} + \frac{b-1}{b}\left[\frac{\partial P}{\partial X}\right]^2 - M_0 C\frac{\partial P}{\partial X}}{\left[1+\frac{b+1}{b}P\right]^2} - (P-P_0)R. \tag{17}$$

Similarly, from (10),

$$\begin{cases} C_p = \dfrac{M_0 CP - (1+2P)\dfrac{\partial P}{\partial X}}{bM_0\left[1+\dfrac{b+1}{b}P\right]} \\ F = \dfrac{M_0 C - \dfrac{b-1}{b}\dfrac{\partial P}{\partial X}}{1+\dfrac{b+1}{b}P}. \end{cases} \tag{18}$$

The expressions for F and C_p possess some interesting features. That for the reduced field, F, is composed of two terms, the first of which expresses Ohm's law, since C is reduced total current density and the denominator is proportional to the local conductivity. The second term is a contribution which is directed away from a hole source, since b is greater than

unity, or since electrons are more mobile than holes. If b were equal to unity, the field would be independent of the concentration gradient. The second term thus represents a departure from Ohm's law which is due to diffusion and which is associated with the presence of current carriers of differing mobilities. It gives a non-vanishing electrostatic field for the case of zero total current. The two terms in the expression for C_p are likewise ohmic and diffusion terms, but here the diffusion term would be present even if the hole and electron mobilities were equal.

Boundary-condition relationships might be illustrated by some examples for this one-dimensional case. If it be specified that for $U > 0$ a fraction f of the total current to the right of a source at the X-origin, say, be carried by holes, then, from (18),

$$\frac{\partial P}{\partial X} = -M_0 \frac{b + (b+1)P}{1 + 2P}\left[f - \frac{P}{b + (b+1)P}\right]C, \qquad X = +0, \quad U > 0. \tag{19}$$

The solution in an X-region to the right of the origin may be determined by this condition and an additional one. The simplest is that for the flow in the semi-infinite region, namely $P = P_0$ for $X = \infty$. This relationship holds for some finite X for an idealized non-rectifying electrode there. For the region between the source and a surface at $X = X_a$ on which there is recombination characterized by a hole transport velocity s, which is also the differential transport velocity for s constant, it is clear that $C = 0$, so that, for $X = X_a$,

$$C_p = -\frac{1}{M_0}\frac{(1+2P)}{b + (b+1)P}\frac{\partial P}{\partial X} = \frac{1}{bM_0}SP; \tag{20}$$

$$S \equiv s/[D_p/\tau]^{\frac{1}{2}}, \qquad s \equiv J_p/p.$$

Consistently with these examples, boundary conditions may in general be expressed as relationships between P, $\frac{\partial P}{\partial X}$, and the parameter C, for given values of X.

A simple transformation of dimensionless quantities serves to extend all of the analytical results which have been given for the n-type semiconductor to the p-type semiconductor: Consider the substitutional transformation which consists in replacing the original dimensional quantities for holes by the corresponding ones for electrons, and vice versa, and in replacing the electrostatic field by its negative. The original set of fundamental equations (1) is invariant under this substitution, which defines an equivalent transformation from the dimensionless quantities of

equations (8) to the desired new set, in which the ratio b of electron to hole mobility is replaced by its reciprocal.

2.32 *The intrinsic semiconductor.*

For the intrinsic semiconductor, in which $p_0 = n_0$, the reduced concentrations given in (8) are inapplicable. As p_0 approaches n_0, these reduced concentrations increase indefinitely, and the equations which those given for the n-type semiconductor approach in the limit are homogeneous in the concentration unit. These limiting equations therefore apply to the intrinsic semiconductor in terms of a concentration unit which may be chosen arbitrarily. The quantity n_0 will be chosen as this unit. Thus, redefining the reduced concentration variables as

$$P \equiv p/n_0\,, \qquad N \equiv n/n_0\,; \qquad P = N, \tag{21}$$

from equations (12) and (14) any two of the equations in the dependent variables P and W given by

$$\left\{\begin{aligned} -(b+1)\ \mathrm{div}\ \mathbf{C}_p &= \frac{2b}{b-1}\ \mathrm{div}\ P\ \mathrm{grad}\ W \\ &= \frac{2b}{b+1}\ \mathrm{div\ grad}\ P = [PQ-1] + \frac{\partial P}{\partial U}, \\ \mathbf{C}_p &= -\frac{1}{b+1}\ P\ \mathrm{grad}\ [W + \log P]; \\ \mathrm{div}\ \mathbf{C} &= 0, \qquad \mathbf{C} = -P\ \mathrm{grad}\left[W - \frac{b-1}{b+1}\log P\right], \end{aligned}\right. \tag{22}$$

and including the right-hand member which is common at least once, characterize the intrinsic semiconductor[24].

It is noteworthy that one of these equations contains only P as dependent variable, W being absent; and this equation indicates that the spatial distribution of carrier concentration is not subject to drift under the field, but only to a diffusion mechanism with diffusion constant $2D_pD_n/(D_p + D_n)$, where $D_n = bD_p$ is the diffusion constant for electrons.[25] This result is readily accounted for as being due to a conductivity in the intrinsic case which is everywhere proportional to the concentration of carriers of either type, so that $\Sigma = P$. The expression for $\mathbf{C}$

[24] These equations for the intrinsic case were first derived quite unambiguously as those for the special case of the parameter p_o/n_o equal to unity in the general equations written in terms of the concentration unit n_o. This unit is, however, less advantageous than $(n_o - p_o)$ which, in obviating much of the formal dependence on p_o, makes for greater generality.

[25] The equations for the intrinsic case might be written in somewhat simpler form by redefining the length unit in terms of $2D_pD_n/(D_p + D_n)$ as a diffusion constant instead of D_p, but their relationship to those of the general case would then be less evident.

in (22) owes its special form simply to this circumstance, while that for $\mathbf{C}_p$ applies also to the general case, and the differential equation in P is a consequence of the equations in P and W from div $\mathbf{C}$ and div $\mathbf{C}_p$. Or, in more detailed terms, since the ohmic contribution to $\mathbf{C}_p$ must be proportional to $\mathbf{C}$, div $\mathbf{C}_p$ contains only the contribution due to diffusion. This is evident from the relationship obtained from (22),

$$\mathbf{C}_p = \frac{1}{b+1}\left[\mathbf{C} - \frac{2b}{b+1}\,\text{grad}\,P\right], \tag{23}$$

from which it follows also that, despite the dependence of the local field on concentration gradient, the ohmic contribution to the hole flow density is the flow density of holes normally present in the intrinsic semiconductor under the unmodulated applied field.

The equations which have been given for one-dimensional flow in the n-type semiconductor can readily be transformed, in the manner indicated, into the corresponding equations for the intrinsic semiconductor.

2.4 *Differential equations in one dimension for the steady state of constant current and properties of their solutions*

The steady state of constant current in one dimension will be considered explicitly for two limiting cases: the n-type semiconductor with $P_0 = 0$, and the intrinsic semiconductor. These serve to illustrate and delimit the qualitative features of the general case. Furthermore, the case $P_0 = 0$ frequently applies as a good approximation[26], as does the intrinsic case, which is of particular interest not only in itself but also because the extrinsic semiconductor exhibits intrinsic behavior for large concentrations, and because moderate increases in temperature above room temperature, such as joule heating may produce, suffice to bring high back voltage germanium into the intrinsic range of conductivity[27]. The temperature dependence of P_0 and of other reduced quantities is evaluated for germanium in the Appendix.

The ordinary differential equations in the reduced hole concentration, P, for the steady state in one dimension, which result from equations (17) and (22) by equating the time derivatives to zero are as follows:

$$\frac{d^2P}{dX^2} = \frac{C\dfrac{dP}{dX} - \dfrac{b-1}{b}\left[\dfrac{dP}{dX}\right]^2}{[1+2P]\left[1+\dfrac{b+1}{b}P\right]} + \frac{P\left[1+\dfrac{b+1}{b}P\right]}{1+2P}R \tag{24}$$

[26] In n-type germanium of resistivity about 5 ohm cm, for example, the electron concentration exceeds the equilibrium hole concentration by a factor of about 70.

[27] Germanium which is substantially intrinsic at room temperature has been produced: R. N. Hall, paper 15 of the Oak Ridge Meeting of the American Physical Society, March 18, 1950.

for the n-type semiconductor with $P_0 = 0$, and

$$\frac{d^2P}{dX^2} = \frac{b+1}{2b}(P-1)R \tag{25}$$

for the intrinsic semiconductor, with R given as $(1 + aP)$ by (16); P has the same meaning in both equations, the concentration unit being n_0 for each case. With time variations excluded in this way, the parameter C is a constant and the differential equations apply to the steady state of constant current.

Since the equations involve only the single independent variable X which does not appear explicitly, their orders may be reduced by one, in accordance with a well-known transformation, which consists in introducing P as a new independent variable, and

$$G \equiv \frac{dP}{dX} \tag{26}$$

as new dependent variable: Noting that $\frac{d}{dX}$ is equivalent to $G\frac{d}{dP}$, the differential equations become

$$\frac{dG}{dP} = \frac{C - \frac{b-1}{b}G}{[1+2P]\left[1+\frac{b+1}{b}P\right]} + \frac{P\left[1+\frac{b+1}{b}P\right]R}{[1+2P]G} \tag{27}$$

for the n-type semiconductor, and

$$\frac{dG}{dP} = \frac{b+1}{2b}\frac{(P-1)R}{G} \tag{28}$$

for the intrinsic semiconductor. These are differential equations of the first order.

The solutions sought in the semi-infinite region, $X > 0$, are those for which $G = 0$ for $\Delta P = 0$, that is, those which pass through the $(\Delta P, G)$ — origin, where ΔP, which denotes $P - P_0$, equals P for the n-type semiconductor and $P - 1$ for the intrinsic semiconductor. This condition is that the concentration gradient vanish with the concentration of added holes, as it must for X infinite. It will be shown that the differential equations possess singular points at the $(\Delta P, G)$-origin, and the physical interpretation of the solutions through these singular points will be examined. For this purpose, consider equation (27) for the n-type semiconductor which, in the neighborhood of the origin, assumes the approximate form,

$$\frac{dG}{dP} = \frac{G}{P} = C + \frac{P}{G}, \tag{29}$$

since R is close to unity for P small, whence

$$\frac{dG}{dP} = \frac{G}{P} = \frac{1}{2}[C \pm \sqrt{C^2 + 4}]. \tag{30}$$

Similarly, for the intrinsic semiconductor, for $P-1$ small,

$$\frac{dG}{d(P-1)} = \frac{G}{P-1} = \pm\sqrt{\frac{b+1}{2b}}. \tag{31}$$

There are thus, in each case, two solutions through the $(\Delta P, G)$-origin, one with a positive derivative and the other with a negative derivative. Consider now the doubly-infinite region with a source at $X = 0$. Then, for $X > 0$, the negative derivatives apply, since the concentration gradient G is negative. Similarly, for $X < 0$, the positive derivatives apply. Now, the value of the current parameter C will be substantially the same in both regions, since it has been assumed that ΔP is small. For C positive, equation (30) for the n-type semiconductor indicates that the magnitude of dG/dP for $X < 0$ exceeds that for $X > 0$, and the situation is reversed if the sign of C is changed. That is, the magnitude of the concentration gradient increases more slowly with concentration for field directed away from a source than for field directed towards a source, which is otherwise plausible. For the intrinsic semiconductor, on the other hand, equation (31) shows that corresponding magnitudes of the concentration gradient are equal and entirely independent of C, a result which the differential equation (28) establishes in general.

It thus appears that a differential equation for the steady state possesses two solutions through the $(\Delta P, G)$-origin, and that one of the solutions corresponds to the case of field directed towards a source, the other to the case of field directed away from a source. Field directed towards a source is called field opposing, while field directed away from a source is called field aiding, the latter being the one commonly dealt with in hole-injection experiments. It should be noted that the cases of field opposing or field aiding can be realized in a given X-region only if it adjoins a semi-infinite region free from sources and sinks. In the region between two sources, neither of these cases applies. L. A. MacColl has shown, through a more detailed consideration of the singularity at the $(\Delta P, G)$-origin, that the two solutions through this point are the only ones through it. The origin is thus a saddle-point of the differential equation, and there exist families of nonintersecting solutions in the $(\Delta P, G)$-plane for which the solutions which intersect at the origin are asymptotes. A solution for an X-region between two sources, for example, is a member of such a family, as is in general any solution determined by boundary conditions at the ends of a finite region in X. Such a solution will be called a solution for a composite case; it approaches asymptotically both a field-opposing and a

field-aiding solution, which is consistent with the qualitative geometry associated with a saddle-point, and with the fact that, in the X-region, a

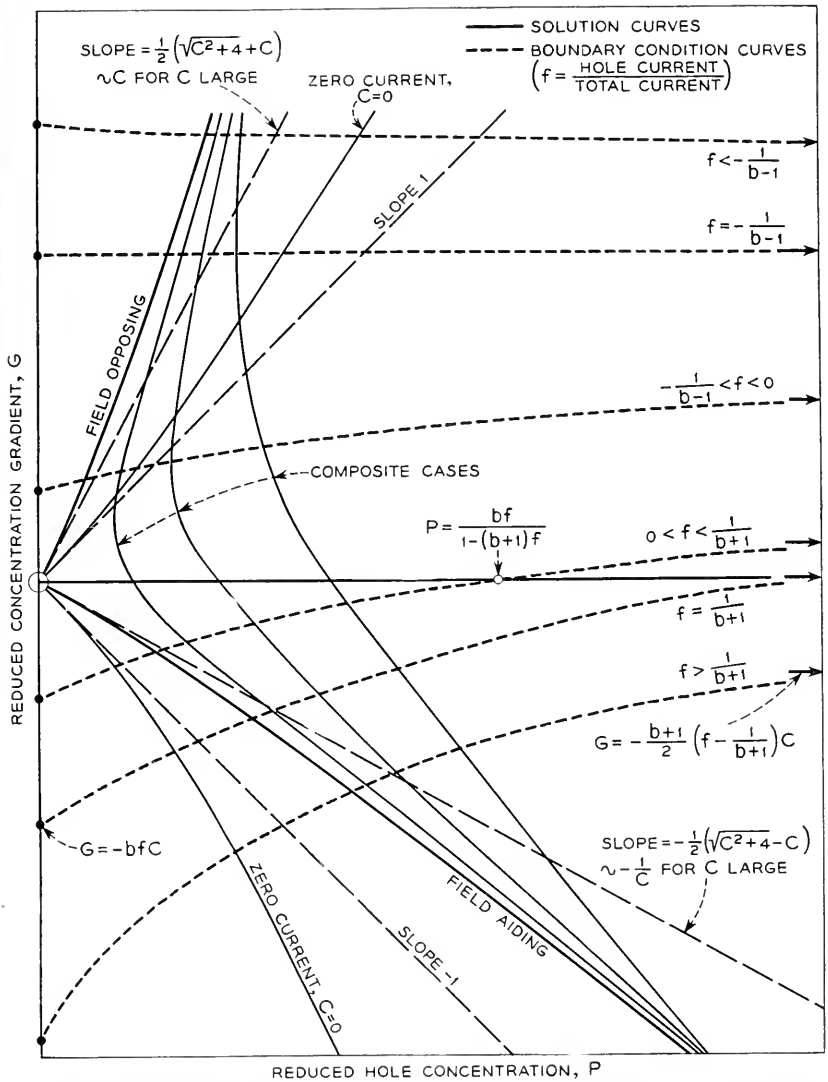

Fig. 1.—Diagrammatic representation in the (P, G)-plane of solutions and boundary conditions for the steady-state one-dimensional flow of holes in an n-type semiconductor.

total current directed away from one source is necessarily directed towards the other. This behavior is illustrated diagramatically for the n-type semiconductor in Fig. 1, which shows, in the (P, G)-plane, solution

curves as well as boundary-condition curves for a source, for a given positive value of C. Those for the intrinsic semiconductor differ only in that the solution curves in the $(P-1, G)$-plane do not depend on C, all being given by the ones for zero total current density, and the corresponding boundary-condition curves are straight lines.

Once a solution, $G(P)$, for field opposing, field aiding, or a composite case, specifying G as a function of P has been obtained, the dependence of P on X is determined by evaluating

$$X = \int_{P^0}^{P} \frac{dP}{G(P)} \tag{32}$$

in accordance with the definition of G, equation (26). For the general composite case, $G(P)$ is that one of the family of solutions for the given C such that the integral between values of P determined by the intersections with the boundary-condition relationships provides the correct interval in X. If P^0 is determined by the condition that for $X = 0$, a fraction f of the total current is carried by holes, then, from (19), P^0 is the point on the solution curve which satisfies either

$$G^0 = -\frac{b + (b+1)P^0}{1 + 2P^0}\left[f - \frac{P^0}{b + (b+1)P^0}\right]C \tag{33}$$

for the n-type semiconductor, or

$$G^0 = -\frac{(b+1)^2}{2b}\left[f - \frac{1}{b+1}\right]C \tag{34}$$

for the intrinsic semiconductor, G^0 being the corresponding value of G.

From the manner of derivation of the boundary conditions (33) and (34), it is evident that they are perfectly general, holding in particular for the cases of field opposing and field aiding, and whatever be the sign of C. The concentration gradient G^0 may be seen to have the correct sign for these cases if it is taken into account that f, defined as C_p/C or I_p/I, may assume any positive or negative value, being positive for field aiding, and negative for field opposing, for which the hole flow is opposite to the applied field. For f negative, the quantities in brackets in equations (33) and (34) are negative. The general principle that the sign of the concentration gradient G is such as to be consistent with the flow of holes from a source requires also that the quantities in brackets be positive for field aiding, or whenever f is positive. For the intrinsic semiconductor, this requires that f for field aiding never be less than $1/(b+1)$. This is clearly a consistent requirement which holds in all generality since, for zero concentration of added holes, or for the normal semiconductor, G^0 vanishes and the ratio of hole current to total current equals $1/(b+1)$.

In the case of the n-type semiconductor, f is not restricted in this way. Consider, for this case, hole injection into the end of a semi-infinite filament, to which the field-aiding solutions apply. As the total current is increased indefinitely, the tangent to the solution in the (P, G)-plane at the origin approaches the P-axis, as does the solution itself, and it is evident from the boundary-condition curves of Fig. 1 that if f is less than $1/(b + 1)$ the hole concentration P^0 at the source approaches as a limit the indicated abscissa of intersection of the appropriate boundary-condition curve with the P-axis, or the value for which the quantity in brackets vanishes. It is similarly evident that P^0 increases indefinitely with total current in either semiconductor if f is greater than or equal to $1/(b + 1)$. This is a result otherwise to be expected from the qualitative consideration that an extrinsic semiconductor becomes increasingly intrinsic in its behavior as the concentration of injected carriers is increased.

Figure 1 serves also to facilitate a count of the number of degrees of freedom which the steady-state solutions possess: Corresponding to values of the concentration and concentration gradient at a point in a semiconductor filament in which added carriers flow, there is a point (P, G) in the half-plane, $P > 0$, of the figure. If the total current density is specified in addition, the value of C and the solution through the point (P, G) are determined. This solution applies in general to a composite case, which therefore possesses three degrees of freedom. That is to say, at a point in a filament, any given magnitudes of both concentration and concentration gradient can be realized for a preassigned total current density by a suitable disposition of sources to the right and left. The cases of field opposing or field aiding, however, possess only two degrees of freedom, since the given concentration and gradient determine the total current density and the solution, which must pass through the origin; and which of the two cases applies depends on whether the point (P, G) lies to the left or to the right of the curves, shown in the figure, for the zero-current solution. Thus, in a filament with a single source of holes, for example, the concentration, concentration gradient, total current density, and any functions of these, such as hole flow density and electrostatic field, are all quantities the specification of any two of which at a point completely determines the solution for a source-free X-region which includes the point.

3. Solutions for the Steady State

For a given value of the current parameter C, solutions for the steady state of constant current in a single cartesian distance coordinate, specifying G in terms of the relative hole concentration P, and P, the reduced hole

flow density C_p, and the reduced electrostatic field F, in terms of reduced distance X are found in general by numerical means, which include numerical integration and the evaluation of appropriate series expansions.

General solutions which have been evaluated numerically for n-type germanium for a number of values of the current parameter are given in the figures. In the limiting cases of P small and P large, analytical approximations for the extrinsic semiconductor are readily obtained, that for P large being derived from an analytical solution for C equal to zero, or zero current. If the steady-state problem for the extrinsic semiconductor is simplified by neglecting either recombination or diffusion, solutions are obtainable which, like the zero-current one, are expressible in closed form.

For the intrinsic semiconductor, the general problem considered in this section is solved quite simply by analytical means. The solution provides, as physical considerations indicate it should, the same analytical approximation for large P as does the zero-current solution for the extrinsic case. It may be well to consider first the intrinsic semiconductor which, aside from the extrinsic semiconductor for the case of zero current, appears to constitute the only analytically solvable steady-state case in one dimension which has physical generality according to the present approach.

3.1 *The intrinsic semiconductor*

Integrating the differential equation (28), it is found that

$$G^2 = \frac{b+1}{b} \int (P-1) R dP, \tag{35}$$

with R given as $1 + aP$ by (16), for an arbitrary combination of the two recombination mechanisms, assumed independent. Thus

$$G^2 = \frac{b+1}{2b} (P-1)^2 \left[(1+a) + \frac{2}{3} a(P-1) \right] \tag{36}$$

for the cases of field opposing or field aiding, for which $G = 0$ for $P-1 = 0$; for a composite case, a suitable constant is included on the right-hand side. Excluding composite cases, the root may be taken in (36) and G replaced by its definition, which gives

$$\frac{d(P-1)}{dX} = \pm \left[\frac{b+1}{2b} \right]^{\frac{1}{2}} [P-1] \left[(1+a) + \frac{2}{3} a(P-1) \right]^{\frac{1}{2}} \tag{37}$$

and if the X-origin is selected more or less arbitrarily as the point at which P is infinite, then (37) gives

$$P - 1 = \frac{3(1+a)}{2a} \operatorname{csch}^2 \left[\frac{(1+a)(b+1)}{8b} \right]^{\frac{1}{2}} X \tag{38}$$

provided $a \pm 0$; for mass-action recombination $a = 1$. For $a = 0$ or for constant mean lifetime, (37) gives an exponential dependence of $P-1$ on X:

$$P - 1 = (P^0 - 1) \exp\left[\pm\left[\frac{b+1}{2b}\right]^{\frac{1}{2}} X\right], \tag{39}$$

where P_0 is the relative hole concentration for $X = 0$. Linear combinations of the two solutions in (39) give solutions for composite cases, since the differential equation from which (39) was derived is linear in P. A similar result does not hold if there is mass-action recombination present, and the more general procedure above referred to must then be followed.

A characteristic feature of these solutions for the intrinsic semiconductor is their independence of the current parameter C, this parameter occurring only through a boundary condition, such as the one given in equation (34) of Section 2.4. They are symmetrical in shape about a source, the dependence of the concentration on the magnitude of the distance from the source being the same for field opposing as for field aiding, which follows quite simply from the symmetrical forms of the solutions, and the condition that the concentration is everywhere continuous.

Equations (22) and (23) of Section 2.32 provide the hole flow density and the electrostatic field for this case. With G given for mass-action recombination or for constant mean lifetime by the appropriate special case of equation (36), and using the positive sign for an X-region to the left of sources and the negative sign for an X-region to the right,

$$\begin{cases} C_p = \dfrac{1}{b+1}\left[C - \dfrac{2b}{b+1} G\right] \\ F = \dfrac{1}{P}\left[C - \dfrac{b-1}{b+1} G\right]. \end{cases} \tag{40}$$

The electrostatic potential, V, is readily expressed in terms of P: From

$$F = -\frac{eL_p}{kT}\frac{dV}{dx} = -\frac{e}{kT}\frac{dV}{dX} = -\frac{e}{kT} G \frac{dV}{dP} \tag{41}$$

and (40), it is found that

$$\frac{e}{kT}\frac{dV}{dP} = \frac{b-1}{b+1}\frac{1}{P} - C\frac{1}{GP}, \tag{42}$$

whence

$$\frac{eV}{kT} = \frac{b-1}{b+1}\log P - C\int\frac{dP}{GP} = \frac{b-1}{b+1}\log P - C\int\frac{dX}{P}, \tag{43}$$

with the integral to be evaluated for the particular case it is desired to consider.

3.2 The extrinsic semiconductor: n-type germanium

The evaluation of steady-state solutions for the extrinsic semiconductor involves, as a first step, the determination of G as a function of P from the differential equation (27), which is accomplished by numerical integration and by the use of series expansions. These variables are subsequently found in terms of X in the manner described in Section 2.4. The series expansions, which are Maclaurin's series in P, and series in powers of the current parameter, C, with coefficients functions of P, are given explicitly for the n-type semiconductor in the Appendix; they readily furnish the corresponding series for the p-type semiconductor by means of the transformation discussed at the end of Section 2.31. The Maclaurin's series in P are useful for starting the solutions at the (P, G)-origin. As P increases, these series converge increasingly slowly, and it becomes necessary to extend the solutions by other means. For the larger values of C, however, the numerical integration for the important case of field aiding becomes increasingly difficult, and it is advantageous to use the appropriate series in the current parameter, which converges the more rapidly the larger is C. The first term alone in this series for field aiding gives in closed form the solution for the case in which diffusion is neglected; and the existence of the series itself was, in fact, originally suggested by the form of the solution for this case[28]. Series of this type are given also for field opposing, and it seems probable that such series are obtainable for composite cases as well, though this has not been investigated.

Solutions were evaluated numerically for n-type germanium, by the means described, using the value 1.5 for the mobility ratio[29], b. For the case of mass-action recombination, solutions for values of the current parameter, C, up to 50, specifying $| G |$ in terms of P, are given in Fig. 2, both for field opposing and field aiding. These solutions in the (P, G)-plane are given to permit the fitting of boundary conditions at a hole source, according to a method described in Section 4. Solutions specifying P in terms of X for field aiding are given in Fig. 3, with the X-origin chosen more or less arbitrarily at $P = 100$. The corresponding solutions for the reduced hole flow density, C_p, and the reduced field, F, are given

[28] The solution for this case was communicated by Conyers Herring and is given in his paper of reference 12.

[29] The hole mobility and the value 1.5 for the mobility ratio were determined by G. L. Pearson from the temperature dependence of the conductivity and Hall coefficient in p-type germanium. J. R. Haynes has recently obtained, from drift-velocity measurements, the same hole mobility, but the larger value 2.1 for the ratio of electron mobility in n-type germanium to hole mobility in p-type: Paper L2 of the Chicago Meeting of the American Physical Society, November 26, 1949.

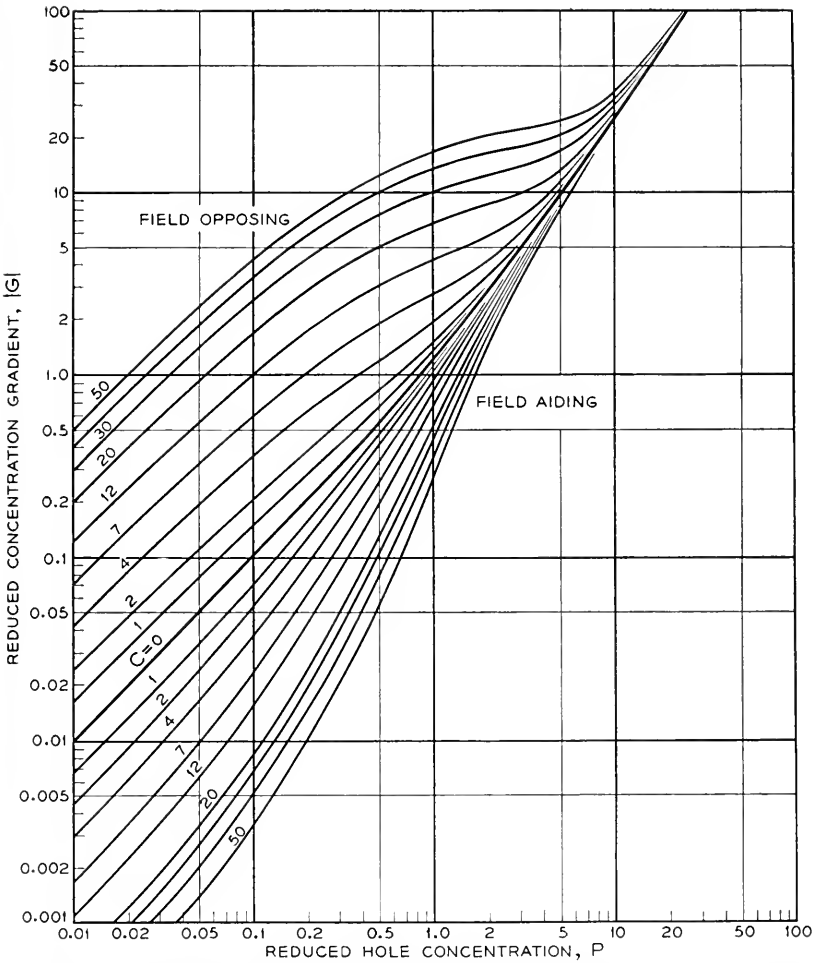

Fig. 2.—The dependence of the reduced concentration gradient on reduced concentration for the steady-state one-dimensional flow of holes with mass-action recombination in n-type germanium.

respectively in Fig. 4 and in Fig. 5. In accordance with equations (10), (18), and (26), the solutions for C_p and F are found from

$$C_p = \frac{1}{b}(FP - G) = \frac{CP - (1 + 2P)G}{b + (b + 1)P}, \tag{44}$$

and

$$F = \frac{C - \frac{b - 1}{b}G}{1 + \frac{b + 1}{b}P}. \tag{45}$$

The electrostatic potential may be evaluated from F in a manner similar to that followed in the preceding section.

3.3 *Detailed properties of the solutions*

The general solutions given in the figures illustrate certain properties which can be established through the analytical approximations obtainable for small and for large values of the relative concentration of added holes. The principal qualitative properties evident from the figures are:

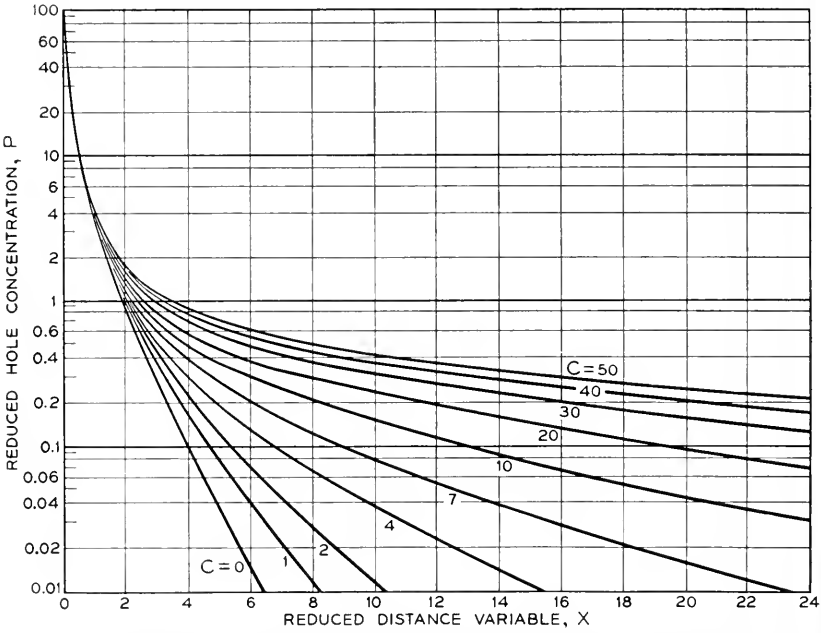

Fig. 3.—The dependence of the reduced concentration on reduced distance for the steady-state one-dimensional flow of holes with mass-action recombination in n-type germanium.

The relative hole concentration, P, and the reduced hole current, C_p, depend exponentially on distance for small concentrations; and for large concentrations all solutions for a given dependent variable run together, independently of the value of the current parameter, and give comparatively rapid variations of hole concentration and current with distance[30]. The property that a common solution independent of total current or

[30] These rapid variations would account for the observation of J. R. Haynes that estimates, for a given emitter current, of hole concentrations or currents in a filament at a point contact removed from the emitter, with no additional applied field, are largely independent of changes in f_e for the emitter.

applied field obtains for large P results from diffusion in conjunction with the increase of conductivity. As may be expected, the solutions for the case of constant mean lifetime also have this property, the recombination law merely affecting the form of the common solution.

In Fig. 6 are shown curves for P, C_p, and F for the case of constant mean lifetime in n-type germanium, evaluated for C equal to 16.3. These curves are intended to illustrate the qualitative differences between the solutions for this case and those for mass-action recombination, which are manifest

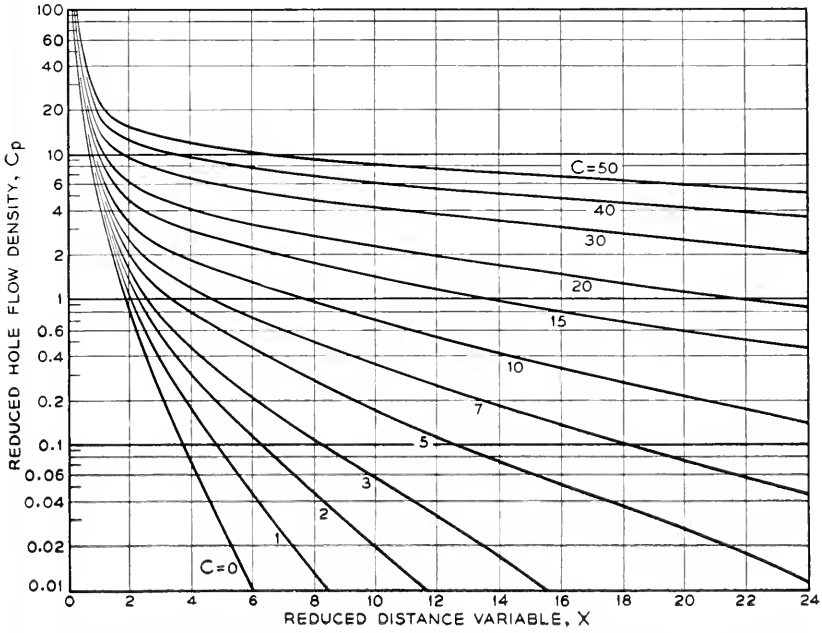

Fig. 4.—The dependence of the reduced hole flow density on reduced distance for the steady-state one-dimensional flow of holes with mass-action recombination in n-type germanium.

primarily at the larger concentrations. The dashed curves in the figure give the corresponding solutions for the case of mass-action recombination; and the X-origins for the two cases have been so chosen that corresponding curves, which exhibit essentially the same dependence on X for small P, coincide in the limit of small P. As the figure shows, constant mean lifetime gives an exponential dependence of P on X for large P, while mass-action recombination gives larger concentration gradients, with an increase of P to indefinitely large values in the neighborhood of a vertical asymptote.

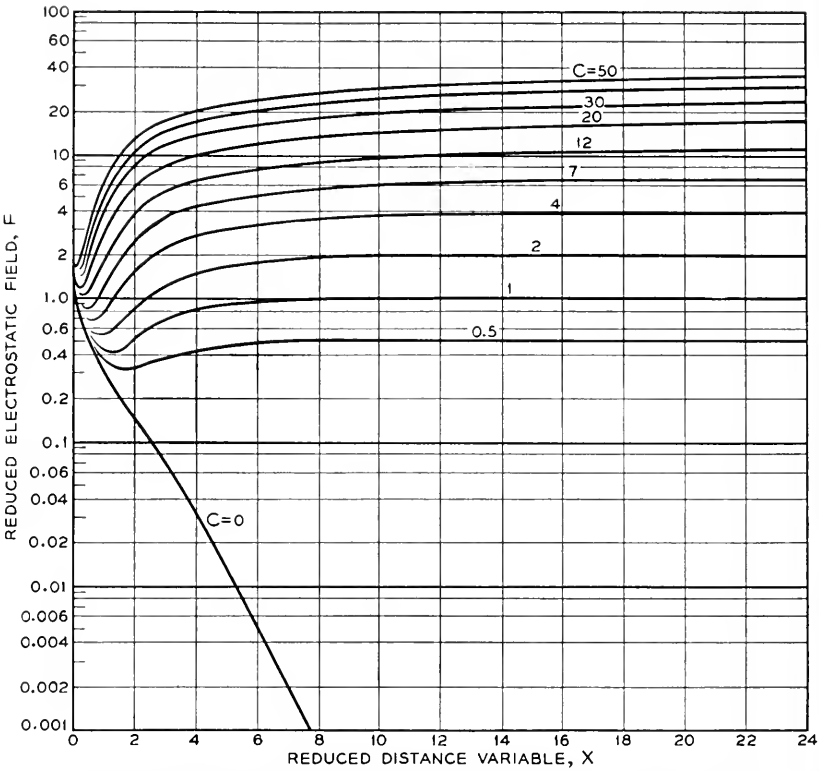

Fig. 5.—The dependence of the reduced electrostatic field on reduced distance for the steady-state one-dimensional flow of holes with mass-action recombination in n-type germanium.

3.31 *The behavior for small concentrations*

The exponential dependence of P and C_p on distance for P small is given for the n-type semiconductor by the analytical approximations,

$$\begin{cases} P = P_s \exp\left[-\tfrac{1}{2}[\pm\sqrt{C^2+4} - C]X\right] \\ C_p = \dfrac{1}{2b}[\pm\sqrt{C^2+4} + C]P, \end{cases} \tag{46}$$

where P_s is a suitable constant. If C is positive, the plus sign holds for field aiding[31] and the minus sign for field opposing. These approximate

[31] It is evident from the curves for C_p in Fig. 4 that the exponential extrapolation back to the emitter location of estimates of hole concentrations or currents at a point contact on a germanium filament lead to values of f_e for the emitter which are too small. Using moderately large injected currents and no additional applied fields, J. R. Haynes once obtained in this way an apparent f_e of about 0.2. From the figure, this is the apparent f_e to be expected for moderate and large values of C for the true f_e equal to unity.

solutions, which hold for any recombination law, are obtained quite simply, by integration, from G in terms of P to the first term of the Maclaurin's expansion, given in the Appendix. It might be noted that for this approximation the electrostatic field is equal to the applied field, so that F equals C.

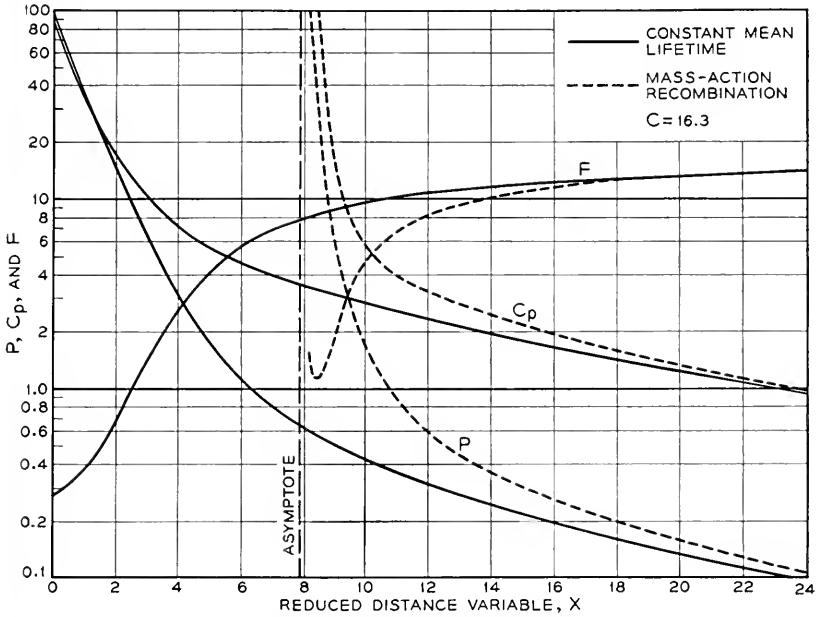

Fig. 6.—The dependence of the reduced hole concentration, hole flow density, and electrostatic field on reduced distance for steady-state one-dimensional hole flow in n-type germanium, for the cases of constant mean lifetime and mass-action recombination.

Since P is small, the transport velocity of holes is equal to their differential transport velocity[32]. Writing the equation for C_p in dimensional form, the transport velocity is found to equal

$$s = \tfrac{1}{2}\left[\pm\sqrt{(\mu E_a)^2 + 4D_p/\tau} + \mu E_a\right], \tag{47}$$

with the plus sign for field aiding and the minus sign for field opposing, if the applied field, E_a, is positive. This result is consistent with the

[32] In accordance with equations (7), (8), and (10), the differential transport velocity for the steady state in one dimension may be found from the general formula,

$$bM_0 dC_p/dP = -(P - P_0)R/G.$$

Its equalling the transport velocity proper for P small appears to result from the property of non-composite cases that the dependent variables, for a given C, are all functions of P which do not depend on any quantity determined by the boundary values, a property which composite cases, with their additional degree of freedom, do not possess.

equation for P, which may be written as

$$P = P_s \exp(-x/s\tau). \tag{48}$$

For a large aiding field, s reduces to the velocity of drift under this field while, for a large opposing field, the magnitude of s is approximately $D_p/\mu E_a \tau$. For zero field, s equals the diffusion velocity $(D_p/\tau)^{\frac{1}{2}}$, which is a diffusion distance for a mean lifetime divided by the mean lifetime. This diffusion velocity can be specified in terms of its field equivalent, or the field which gives an equal drift velocity, and for germanium it is found that the equivalent field is about 8 volt cm^{-1} for τ equal to one microsecond and about 2.5 volt cm^{-1} for τ equal to 10 microseconds.

For small concentrations of added holes in the intrinsic semiconductor, or $(P-1) << 1$, equations (38) and (40) give the approximate solutions,

$$\begin{cases} P - 1 = (P^0 - 1) \exp\left[\pm\left[\dfrac{(1+a)(b+1)}{2b}\right]^{\frac{1}{2}} X\right] \\ C_p = \dfrac{1}{b+1}\left[C - \dfrac{2b}{b+1} G\right] \sim \dfrac{1}{b+1} C, \end{cases} \tag{49}$$

the X-origin being selected arbitrarily at the point at which the relative concentration is P^0 according to the approximation. It is evident from the equation for C_p that, for $(P-1)$ small, the transport velocity is the drift velocity under the applied field, which is the velocity of the holes normally present in the semiconductor. The differential transport velocity, obtainable by differentiating the equation for C_p with respect to P and using the differential equation (28), or by writing the exponent in the equation for $(P-1)$ in the form given in (48), is, on the other hand, given by

$$s = \left[\frac{2b}{(1+a)(b+1)}\right]^{\frac{1}{2}} \left[\frac{D_p}{\tau}\right]^{\frac{1}{2}} = \left[\frac{1}{1+a} \frac{2D_p D_n}{D_p + D_n} \Big/ \tau\right]^{\frac{1}{2}}, \tag{50}$$

and is a diffusion velocity. This holds for holes added in any concentration if $a = 0$, or for constant mean lifetime, since the first of equations (49) is then the general solution given in (39).

The nature of the flow for small concentrations of added carriers in the general case, which depends on the parameter P_0, is illustrated qualitatively by the n-type and intrinsic cases considered, for which P_0 is respectively zero and infinite. Solutions for the general case are easily evaluated analytically from the linear differential equation which results from (17) if $P - P_0 << \frac{1}{2} + P_0$. It can be shown from the field-aiding steady-state solution that the ratio of the differential transport velocity to the velocity, proportional to C, of drift under the applied field is for $C^2 >> (1 + 2P_0)M_0$ equal to the quantity $1/M_0$. This result is consistent

with those already derived: For large applied aiding fields, the differential transport velocity changes from the drift velocity, for P_0 equal to zero and M_0 unity, to the diffusion velocity given in (50) as P_0 and M_0 increase indefinitely.

3.32 *The zero-current solutions and the behavior for large concentrations*

The solutions for the intrinsic semiconductor for the current parameter equal to zero are, of course, the same as the general ones given in Section 3.1, since the current parameter does not occur in the differential equation. For the n-type semiconductor, the differential equation (27) becomes an equation of the Bernouilli type for C equal to zero, and may be solved by quadratures. It is then linear in G^2, and gives, for field aiding or field opposing,

$$G^2 = 2\left[\frac{1 + \frac{b+1}{b}P}{1+2P}\right]^2 \int_0^P \frac{P(1+P)(1+aP)}{1+\frac{b+1}{b}P}\,dP, \tag{51}$$

expressing the recombination function R according to equation (16) for a combination of the two recombination mechanisms. Writing, for brevity,

$$\begin{cases} \beta \equiv \dfrac{b}{b+1} \\ M \equiv 1 + \dfrac{b+1}{b}P, \end{cases} \tag{52}$$

and evaluating the integral in (51), the following result is obtained:

$$\begin{aligned} G^2 = 2\beta^2\left[\frac{M}{1+2P}\right]^2 \Big[& [1-a] \\ & \cdot [\beta(M^2-1) + (1-4\beta)(M-1) - (1-2\beta)\log M] \\ & + a[\tfrac{2}{3}\beta^2(M^3-1) + \tfrac{3}{2}(1-2\beta)(M^2-1) \\ & + (1-6\beta+6\beta^2)(M-1) - (1-\beta)(1-2\beta)\log M]\Big]. \end{aligned} \tag{53}$$

For P large, this solution gives the approximations,

$$G = \pm\left[\frac{b+1}{2b}\right]^{\frac{1}{2}} P \tag{54}$$

for constant mean lifetime, with $a = 0$, and

$$G = \pm\left[\frac{a(b+1)}{3b}\right]^{\frac{1}{2}} P^{\frac{3}{2}} \tag{55}$$

if there is mass-action recombination present, so that $a \neq 0$. The dependence of P on X for these approximations is readily obtained by integrating the differential equations which result from writing in place of G, its definition, dP/dX; constant mean lifetime gives an exponential dependence. An examination of (54) and (55) in conjunction with the general differential equation (27) shows that, for P large, the dominant term in the differential equation is independent of C. It follows that solutions for all values of C approach a common solution for P large, which is given by (54) or (55). The solutions run together appreciably for P sufficiently large that P and M are substantially proportional, that is, for P large compared with $b/(b + 1)$, which is of order unity. It is to be expected that the approximations (54) and (55) should apply equally well to the intrinsic semiconductor, and this expectation is easily verified by evaluating the integral in equation (35) for the intrinsic semiconductor, for P large, for the two recombination cases here considered.

4. Solutions of Simple Boundary-Value Problems for a Single Source

Among the boundary-value problems whose solutions are useful in the interpretation of data from experiments in hole injection are the following: the semi-infinite filament for field aiding, with holes injected at the end, which constitutes a relatively simple case; and the doubly-infinite filament with a single plane source, with which this section will be primarily concerned.

Consider first the semi-infinite filament, and suppose that it starts at the X-origin and extends over positive X, so that the current parameter is positive for field aiding. If two quantities are specified, namely the current parameter and the fraction f_ϵ of the current carried by holes at the origin or injection point, then the solution of the boundary-value problem is completely determined. It is merely necessary to select the general field-aiding solution for P or C_p in terms of X, for the particular value of the current parameter, and then to determine the X-origin, corresponding to the source, which is simply the X at which the ratio f of C_p to C equals f_ϵ.

Use in the boundary-condition equations (33) and (34) of the approximate expressions given in (54) and (55) for G in terms of P, for large P, permits the complete analytical determination of the dependence of P^0 on total current as this current is indefinitely increased. It was shown in Section 2.4 that, if f_ϵ is less than $1/(b + 1)$ for the n-type semiconductor, P^0 approaches as a limit the value for which G^0 vanishes according to the boundary-condition equation (33); in all other cases for the n-type semiconductor, or if f_ϵ exceeds $1/(b + 1)$ for the intrinsic semiconductor, P^0 increases indefinitely with C. For $f_\epsilon > 1/(b + 1)$, it is readily seen that

P^0 is proportional in the limit to C for constant mean lifetime, and to $C^{\frac{2}{3}}$ for mass-action recombination; and, for $f_\epsilon = 1/(b+1)$ in the case of the n-type semiconductor, P^0 increases as $C^{\frac{1}{3}}$ for constant mean lifetime, and as $C^{\frac{2}{5}}$ for mass-action recombination.

Consider now the doubly-infinite semiconductor filament with a source at the origin, and suppose that the total injected current at the source is C_ϵ, in reduced form, with a fraction f_ϵ of this current carried by holes. Denote by C^- and by C^+ the reduced total currents for $X < 0$ and for $X > 0$, respectively. Since the injection of holes requires that C_ϵ be positive, at least one of C^- and C^+ must be positive, since total current is conserved. Let f^- and f^+ denote, respectively, the ratio of the hole current at the origin to the left, C_p^-, to the total current C^-, and the ratio of the hole current at the origin to the right, C_p^+, to the total current, C^+. It might be noted that, for a flow of holes to the left, say, against the field, C^- and C^+ are positive and f^- is negative, and that, if C^- is (plus) zero, f^- is (negatively) infinite, corresponding to the flow of holes under zero applied field. Now, general boundary-condition equations of the form of (33) or (34) hold with the sign conventions here employed, as indicated in Section 2.4. One may be written for the flow to the left, another for the flow to the right, making use of the condition that the relative concentration P is everywhere continuous; G exhibits a discontinuity of the first kind at the source, with a change in sign. Writing G^- for the limiting value of the reduced concentration gradient as the origin is approached from the left, and G^+ the limiting value as the origin is approached from the right, the boundary-condition equations are, for the n-type semiconductor,

$$\text{(56)} \qquad \begin{cases} G^- = -\dfrac{b + (b+1)P^0}{1 + 2P^0}\left[f^- - \dfrac{P^0}{b + (b+1)P^0}\right]C^- \\ G^+ = -\dfrac{b + (b+1)P^0}{1 + 2P^0}\left[f^+ - \dfrac{P^0}{b + (b+1)P^0}\right]C^+. \end{cases}$$

For the intrinsic semiconductor, they are

$$\text{(57)} \qquad \begin{cases} G^- = -\dfrac{(b+1)^2}{2b}\left[f^- - \dfrac{1}{b+1}\right]C^- \\ G^+ = -\dfrac{(b+1)^2}{2b}\left[f^+ - \dfrac{1}{b+1}\right]C^+. \end{cases}$$

There are, in addition, an equation which expresses the conservation of hole flow, and one which expresses the conservation of total current, as follows:

$$\text{(58)} \qquad \begin{cases} f^+ C^+ - f^- C^- = f_\epsilon C_\epsilon \\ C^+ - C^+ = C_\epsilon \quad . \end{cases}$$

The solution of the problem is determined by f_ϵ and the three parameters which specify the total currents: With these four quantities known, then, from equations (56) or (57) in conjunction with (58) and the known general solutions in the $(\Delta P, G)$-plane which apply to the left and to the right of the origin, all of the quantities P^0, G^-, G^+, f^- and f^+ can be found and the problem completely solved.

The technique of obtaining the solution depends on a simple fundamental result which may be expressed as follows:

For fixed f_ϵ and C_ϵ, consider the sum of the magnitudes of the concentration gradients at a single common source from which holes flow into a number of similar filaments in parallel, for any consistent distribution among the filaments of total currents, some of which may be produced by opposing fields. This sum is equal to the magnitude of the concentration gradient at the source if the entire flow, under the appropriate aiding field, were confined to a single filament.

The total magnitude of the concentration gradient, in this sense, is an invariant for fixed f_ϵ and C_ϵ. Specifically, for the n-type semiconductor, it follows from equations (56) and (58) that

$$G^+ - G^- = -\frac{b + (b+1)P^0}{1 + 2P^0}\left[f_\epsilon - \frac{P^0}{b + (b+1)P^0}\right]C_\epsilon. \tag{59}$$

Similarly, for the intrinsic semiconductor,

$$G^+ - G^- = -\frac{(b+1)^2}{2b}\left[f_\epsilon - \frac{1}{b+1}\right]C_\epsilon. \tag{60}$$

The left-hand sides of these equations are the negative of the sum of the magnitudes of the reduced concentration gradients, since G^- is always positive and G^+ always negative, and their right-hand sides are similar in form to those of equations (56) and (57), with the quantities f_ϵ and C_ϵ, characteristic of the source, replacing f^- and C^-, or f^+ and C^+.

The particular utility of these equations arises from their independence of the unknowns f^- and f^+. By means of equation (59) for the n-type semiconductor the evaluation of the five unknown quantities can now be effected as follows: With the current parameters known, the solutions in the (P, G)-plane to the left and right of the X-origin are determined; either both solutions are for field aiding, or else one is for field aiding and the other for field opposing. From them, the sum of the magnitudes of the reduced concentration gradients can be found as a function of P. It is also given, for the origin, as a function of the unknown P^0, by equation (59). The values of the sum for the origin and of P^0 are accordingly found as those which satisfy both relationships. The value of P^0 thus found determines both G^- and G^+ from the respective solutions

in the (P, G)-plane, and f^- and f^+ may be obtained by solving for them in equations (56).

For the intrinsic semiconductor, this method can be applied analytically, and the solution so obtained serves at the same time as an approximation for large relative hole concentrations in the n-type semiconductor, for which the method is otherwise essentially graphical or numerical in the general case. Making use of the symmetry of the solutions for the intrinsic semiconductor about a source, it follows from (57) and (58) that

$$\begin{aligned} G^+ = -G^- &= -\frac{(b+1)^2}{4b}\left[f_\epsilon - \frac{1}{b+1}\right]C_\epsilon \\ &= -\frac{(b+1)^2}{2b}\left[f^+ - \frac{1}{b+1}\right]C^+ = \frac{(b+1)^2}{2b}\left[f^- - \frac{1}{b+1}\right]C^-, \end{aligned} \tag{61}$$

whence

$$\begin{cases} f^- = \dfrac{1}{b+1} - \dfrac{1}{2}\left[f_\epsilon - \dfrac{1}{b+1}\right]\dfrac{C_\epsilon}{C^-} \\ f^+ = \dfrac{1}{b+1} + \dfrac{1}{2}\left[f_\epsilon - \dfrac{1}{b+1}\right]\dfrac{C_\epsilon}{C^+}. \end{cases} \tag{62}$$

It is easily verified that this result holds approximately for large relative concentrations in the n-type semiconductor. Three simple special cases of (62) might be considered: The first is

$$\begin{cases} C^- = -C^+ = -\tfrac{1}{2}C_\epsilon \\ f^- = f^+ = f_\epsilon . \end{cases} \tag{63}$$

This is the rather trivial case of symmetrical flows from a source which supplies all currents. A second special case is that for which C^- and C^+ are both positive, say, and such that there is no hole flow to the left against the field. It is readily found that, for this case,

$$\begin{cases} C^- = \dfrac{b+1}{2}\left[f_\epsilon - \dfrac{1}{b+1}\right]C_\epsilon ; \quad C^+ = \dfrac{b+1}{2}\left[f_\epsilon + \dfrac{1}{b+1}\right]C_\epsilon \\ f^- = 0; \quad f^+ = \dfrac{2}{b+1}\dfrac{f_\epsilon}{f_\epsilon + 1/(b+1)}. \end{cases} \tag{64}$$

Here, the drift from the left under the applied field of holes normally present in the intrinsic semiconductor just cancels the diffusion from the source to the left.[33] A third special case is that in which the total current

[33] Using the numerically obtained solutions, the validity of (64) as an approximation for large concentrations in n-type germanium may be seen as follows: For f_ϵ equal to unity and C_ϵ, C^- and C^+ equal to 2, 1.5, and 3.5 respectively, P^0 is about 0.6 and the fraction of injected holes which flows against the field is nearly one-half; doubling these current densities increases P^0 to 1.45 and decreases the fraction to about one-fourth, and the fraction is less than about one-tenth if the current densities are increased so that C^+ exceeds 15

to the left of the source is zero, the left-hand side of the filament being open-circuited. For this case, equations (62) are better written in the form obtained by multiplying through by C^- or C^+, and the special case in question is then found to be given by

$$(65)\quad \begin{cases} C^- = 0; \qquad C^+ = C_\epsilon \\ C_p^- = -\frac{1}{2}\left[f_\epsilon - \frac{1}{b+1}\right] C_\epsilon; \quad C_p^+ = \frac{1}{2}\left[f_\epsilon + \frac{1}{b+1}\right] C_\epsilon, \end{cases}$$

according to which, if f_ϵ is equal to unity, the magnitude of the hole flow to the left into the open-circuit end is $b/(b+2)$ times that into the circuit end, to the right; or a fraction $b/2(b+1)$ of the holes flows to the left, and a fraction $(b+2)/2(b+1)$ to the right. Thus, for germanium, the hole flow into the open-circuit end is 0.43 as large as that into the circuit end, a fraction 0.30 flowing to the left, and 0.70 to the right. It might be observed that the fractions of the injected holes which flow to the left and right are, in this case, proportional to the total currents C^- and C^+ of the preceding case, for which there is zero hole flow to the left.

Another general limiting case for the n-type semiconductor is that for P_0 small, so that the exponential approximations of Section 3.31 apply. The restriction on the magnitude of P is $P << \frac{1}{2}$. This restriction obtains if C_ϵ is sufficiently small that C^- and C^+ do not differ appreciably. Equation (59) then gives

$$(66)\qquad G^+ - G^- = -bf_\epsilon C_\epsilon .$$

Writing C for C^- and C^+, equations (30) and (46) result in

$$(67)\quad \begin{cases} G^- = \frac{1}{2}\left[\sqrt{C^2+4} + C\right] P^0 = bC_p^+ \\ G^+ = -\frac{1}{2}\left[\sqrt{C^2+4} - C\right] P^0 = bC_p^-, \end{cases}$$

whence, solving for $G^+ - G^-$ and comparing with equation (66),

$$(68)\qquad P^0 = bf_\epsilon C_\epsilon/\sqrt{C^2+4}.$$

In accordance with (67), then,

$$(69)\quad \begin{cases} C_p^- = -\frac{1}{2}\left[1 - C/\sqrt{C^2+4}\right] f_\epsilon C_\epsilon \\ C_p^+ = \frac{1}{2}\left[1 + C/\sqrt{C^2+4}\right] f_\epsilon C_\epsilon. \end{cases}$$

These are the reduced hole flows to the left and right of the source. While it has been assumed that C_ϵ is small compared with C, no restriction has been placed on C itself. For C small compared with unity, the equations indicate that the hole flows to the left and right are the

same in magnitude, while for C large compared with unity,

$$\begin{cases} C_p^- \sim -\dfrac{1}{C^2} f_\epsilon C_\epsilon \\ C_p^+ \backsim f_\epsilon C_\epsilon . \end{cases} \tag{70}$$

Thus, according to this approximation, C should exceed about 10 if no more than one per cent of the holes are to flow against the field. From (75) in the Appendix, a value of 10 for C corresponds to a current density of about 1.2 amp cm^{-2} in germanium of 10 ohm cm resistivity, with τ equal to 10 μsec. This current density is moderately large among those which have been employed in experiments with germanium filaments.

Experimentally, the ideal one-dimensional geometry postulated in the present treatment of the problem of the single source in an infinite filament cannot easily be realized, hole injection generally being accomplished through a point contact or a side arm on one side of the actual filament. If suitable averages are employed, non-uniformity in P at the injection cross-section does not, however, vitiate the approximate results for ΔP large and ΔP small, since their applicability depends largely on the validity over the injection cross-section of the approximation assumed.

Acknowledgment

The author is indebted to a number of his colleagues for their stimulating interest and encouragement; to J. Bardeen and W. Shockley for a number of valuable and helpful comments, as well as to W. H. Brattain, J. R. Haynes, C. Herring, L. A. MacColl, G. L. Pearson, and R. C. Prim. J. Bardeen also suggested the numerical analysis for n-type germanium which constituted one of the initial points of attack, and aided materially in its inception. The rather difficult numerical integrations and associated problems were ably handled by R. W. Hamming, Mrs. G. V. Smith and J. W. Tukey.

5. APPENDIX

5.1 *The concentrations of ionized donors and acceptors*

While the donor and acceptor concentrations need not, of course, be considered for the intrinsic semiconductor, for the extrinsic semiconductor the fundamental equations, as they have been written, are in principle incomplete: Two additional equations in the variables D^+ and A^- are required. One of the required equations is trivial, since changes in the concentration of ionized centers which are compensated by those which determine the conductivity type of the extrinsic semiconductor

can certainly be neglected. For an n-type semiconductor, for example, the term $(A^- - A_0^-)$ in Poisson's equation may be suppressed. This procedure is strictly consistent with the neglect of p_0 and g_0, but undoubtedly holds to an even better approximation. If D is the total donor concentration in the n-type semiconductor, the concentration of ionized donors may be considered to satisfy the equation,

$$\frac{\partial D^+}{\partial t} = H(D - D^+) - KD^+n, \tag{71}$$

which applies to the homogeneous semiconductor, with H and K constants which characterize, respectively, the rate of ionization of unionized donors, and the rate of recombination of an ionized donor with an electron. If, as a result of a small thermal ionization energy, most of the donors are ionized, so that $KD/H << 1$, the change in ionized-donor concentration for the steady state is given by (71) as

$$D^+ - D_0^+ \sim -\frac{KD}{H}(n - n_0), \tag{72}$$

which is small compared with the corresponding change in electron concentration. In other cases, the use of the general expression obtainable from (71) for the steady-state concentration of ionized donors in terms of the electron concentration, or the expression for the other limiting case of relatively few ionized donors, might provide a more precise description provided the conditions under which solutions are sought do not involve unduly rapid changes with time.

5.2 *The carrier concentrations at thermal equilibrium*

The ratio of the thermal-equilibrium values of the hole and electron concentrations may be evaluated for n-type germanium from[5]

$$\begin{cases} np = 3 \cdot 10^{32} T^3 \exp\left(-\frac{8700}{T}\right) \equiv n_i^2 \\ n - p = n_s \sim n_0 = 1/b\mu e\rho_0 = 2.40 \cdot 10^{15}/\rho_0 , \end{cases} \tag{73}$$

where the electron concentration excess n_s corresponds to complete ionization of the donors, and is approximately n_0 at the highest temperature at which P_0 is still negligible, which may be taken as room temperature[29]. The resistivity ρ_0 is that which determines n_0. Thus,

$$P_0 = \tfrac{1}{2}\left[\sqrt{1 + 4(n_i/n_0)^2} - 1\right], \tag{74}$$

[29] loc. cit.
[5] loc. cit.

with n_i, the concentration of holes or electrons in intrinsic germanium at T deg abs, given in (73). It may be estimated that temperature rises of less than 100 deg C will make 10 ohm cm n-type germanium substantially intrinsic in its behavior.

The range of values of the parameter C for which the numerical solutions are given corresponds, for example, to current densities up to the order of 10 amp cm^{-2} in germanium filaments of about 10 ohm cm resistivity, for the mean lifetime τ about 10 μsec; for this mean lifetime, the distance unit L_p is approximately $2 \cdot 10^{-2}$ cm. Current densities corresponding to the larger values of C will ordinarily produce appreciable joule heating in filaments some 10^{-3} cm^2 in area of cross-section, cemented to a backing, with temperature rises of the order of 100 deg C.

The effect of joule heating on L_p and C may be evaluated from

$$(75) \qquad \begin{cases} L_p = 6.6 \left[\dfrac{T}{300}\right]^{-\frac{1}{4}} \tau^{\frac{1}{2}} \\ C = 2.6 \cdot 10^2 \left[\dfrac{T}{300}\right]^{-\frac{5}{4}} \tau^{\frac{1}{2}} \rho\, I, \end{cases}$$

where τ is expressed in sec, I in amp cm^{-2}, and ρ is the normal resistivity in ohm cm of the germanium at T deg abs. These are obtained from the definitions (8), taking the hole mobility in the thermal scattering range to be proportional to $T^{-\frac{3}{2}}$, with the value 1700 cm^2 volt^{-1} sec^{-1} at 300 deg abs.[7]

5.3 *Series solutions for the extrinsic semiconductor in the steady state*

Maclaurin's series for G in the relative concentration P are of the form

$$(76) \qquad G = a_1 P + a_2 P^2 + a_3 P^3 + \ldots$$

for the cases of field opposing and field aiding, the solutions passing through the (P, G)-origin. Substituting the series (76) for G in the differential equation (27) for the n-type semiconductor in the steady state, it is found, in accordance with (30), that

$$(77) \qquad a_1 = \tfrac{1}{2} \left[C \pm \sqrt{C^2 + 4}\right],$$

the sign of C being taken before the radical for field opposing, the other sign for field aiding. The other coefficients are given in terms of a_1 and

[7] loc. cit.

also the b, C, and the constant, a, of the recombination function:

$$
(78)\quad \begin{cases} a_2 = \dfrac{4a_1^2 - \left[2\dfrac{b+1}{b} + a\right]}{C - 3a_1} \\[2ex] a_3 = \dfrac{2a_2^2 + \dfrac{11b+1}{b}a_1a_2 + 2\dfrac{b+1}{b}a_1^2 - \dfrac{b+1}{b}\left[\dfrac{b+1}{b} + 2a\right]}{C - 4a_1} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{cases}
$$

The series in the current parameter are series in ascending powers of the reciprocal of C. Writing, for convenience,

$$
(79)\qquad \gamma \equiv 1/C,
$$

the differential equation (27) may be put in the form,

$$
(80)\qquad \gamma[1 + 2P]\left[1 + \frac{b+1}{b}P\right]GG' + \gamma\frac{b-1}{b}G^2 - G - \gamma P\left[1 + \frac{b+1}{b}P\right]^2 R = 0,
$$

using the prime to denote differentiation with respect to P. Consider expansions of the form,

$$
(81)\qquad G = \sum_{j=j_0}^{\infty} A_j\gamma^j,
$$

in which the A's are functions of P to be determined. Substituting in the differential equation, there results

$$
(82)\qquad \sum_{j=j_0}^{\infty}\sum_{m=j_0}^{\infty}\left[[1 + 2P]\left[1 + \frac{b+1}{b}P\right]A_jA_m' + \frac{b-1}{b}A_jA_m\right]\gamma^{j+m+1} - \sum_{j=j_0}^{\infty} A_j\gamma^j - P\left[1 + \frac{b+1}{b}P\right]^2 R\gamma = 0.
$$

Since the expansions are to hold for arbitrary values of γ, the A's must, for the cases of field opposing and field aiding, for which the solutions pass through the (P, G)-origin, vanish identically for P equal to zero, and be determined by equating to zero the coefficients of given powers of γ in (82). It can, without loss of generality, be assumed that the coefficient of the leading term in the expansion, A_{j_0}, is not identically zero. Then, from (82), it is found that there is no expansion for $j_0 = 0$, that is, no expansion starting with a term independent of γ. Formal expansions can be obtained, however, for $j_0 = -1$ and for $j_0 = +1$. These may be identified,

respectively, with the solutions for field opposing and field aiding, as will be seen.

For $j_0 = -1$, or field opposing, (82) leads to differential equations of the first order for the determination of the A's. The condition that these functions vanish identically for $P = 0$ suppresses all A's of even order. The first term of the expansion is found by solving

$$(83)\quad A'_{-1} + \frac{b-1}{b}\frac{A_{-1}}{[1+2P]\left[1+\frac{b+1}{b}P\right]} = \frac{1}{[1+2P]\left[1+\frac{b+1}{b}P\right]}$$

whence

$$(84)\quad A_{-1} = \frac{P}{1+2P}.$$

The second term is found from

$$(85)\quad A'_1 + \frac{b-1}{b}\frac{A_1}{[1+2P]\left[1+\frac{b+1}{b}P\right]} = \left[1+\frac{b+1}{b}P\right]R,$$

whence, with R equal to unity and $(1+P)$, respectively,

$$(86)\quad \begin{cases} A_1 = \dfrac{P[1+P]\left[1+\frac{b+1}{b}P\right]}{1+2P} & \text{for constant mean lifetime} \\ A_1 = \dfrac{P\left[1+\frac{3}{2}P+\frac{2}{3}P^2\right]\left[1+\frac{b+1}{b}P\right]}{1+2P} & \text{for mass-action recombination.} \end{cases}$$

For the third term, making use of (84), (85) and (86),

$$(87)\quad \begin{cases} A'_3 + \dfrac{b-1}{b}\dfrac{A_3}{[1+2P]\left[1+\frac{b+1}{b}P\right]} \\ \qquad = -[1+P]\left[1+\dfrac{b+1}{b}P\right]^2 \text{ for constant mean lifetime} \\ A'_3 + \dfrac{b-1}{b}\dfrac{A_3}{[1+2P]\left[1+\frac{b+1}{b}P\right]} \\ = -[1+P]\left[1+\frac{3}{2}P+\frac{2}{3}P^2\right]\left[1+\dfrac{b+1}{b}P\right]^2 \text{ for mass-action recombination} \end{cases}$$

whence

$$
(88)\quad \begin{cases}
A_3 = \dfrac{-P\left[1 + \dfrac{4b+1}{2b} P + \dfrac{5b+3}{3b} P^2 + \dfrac{b+1}{2} P^3\right]\left[1 + \dfrac{b+1}{b} P\right]}{1 + 2P} & \text{for constant mean lifetime} \\[2ex]
A_3 = \dfrac{-P\left[1 + \dfrac{33b+6}{12b} P + \dfrac{70b+27}{18b} P^2 + \dfrac{73b+43}{24b} P^3 + \dfrac{38b+30}{30b} P^4 + \dfrac{2b+2}{9b} P^5\right]\left[1 + \dfrac{b+1}{b} P\right]}{1 + 2P} & \text{for mass-action recombination.}
\end{cases}
$$

For $j_0 = +1$, or field aiding, the A's are determined somewhat more simply, recursive relationships obtaining. The results are:

$$
(89)\quad A_1 = -P\left[1 + \frac{b+1}{b} P\right]^2 R,
$$

and

$$
(90)\quad \begin{cases}
A_3 = [1 + 2P]\left[1 + \dfrac{b+1}{b} P\right] A_1 A_1' + \dfrac{b-1}{b} A_1^2 \\[2ex]
A_5 = [1 + 2P]\left[1 + \dfrac{b+1}{b} P\right] [A_1 A_3]' + 2\,\dfrac{b-1}{b} A_1 A_3 \\[2ex]
A_7 = [1 + 2P]\left[1 + \dfrac{b+1}{b} P\right] [[A_1 A_5]' + A_3 A_3'] \\
\qquad + 2\,\dfrac{b-1}{b}\left[A_1 A_5 + \dfrac{1}{2} A_3^2\right] \\[2ex]
A_9 = [1 + 2P]\left[1 + \dfrac{b+1}{b} P\right] [[A_1 A_7]' + [A_3 A_5]'] \\
\qquad + 2\,\dfrac{b-1}{b} [A_1 A_7 + A_3 A_5] \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{cases}
$$

The identification of the series in the parameter γ as series for field opposing and field aiding is accomplished by evaluating them for small P and then comparing them with the first terms of the corresponding Maclaurin's series in P, expanded in powers of γ. Further agreement is

obtained by comparing the first terms of the series in γ with the functions of P which result from evaluating the Maclaurin's series for γ small.

5.4 *Symbols for Quantities*

$a \equiv \tau/\tau_v$, constant in recombination function.
$a_j \equiv$ coefficients in the Maclaurin's expansion of G in powers of P; j an integer.
$A_j \equiv$ coefficients in the expansion of G in powers of γ; j an integer.
$A^- \equiv$ concentration of ionized acceptors.
$A_0^- \equiv$ thermal-equilibrium concentration of ionized acceptors.
$b \equiv$ ratio of electron mobility to hole mobility.
$C \equiv I/I_0$, reduced total current density.
$C_\epsilon \equiv$ reduced emitter current.
$C^- \equiv$ reduced total current to the origin from the left.
$C^+ \equiv$ reduced total current from the origin to the right.
$C_n \equiv -I_n/I_0$, reduced electron flow density.
$C_p \equiv I_p/I_0$, reduced hole flow density.
$\gamma \equiv I/C$.
$\Gamma \equiv \epsilon/4\pi\sigma_0\tau$, reduced time for the dielectric relaxation of charge.
$D \equiv$ total donor concentration.
$D^+ \equiv$ concentration of ionized donors.
$D_0^+ \equiv$ thermal-equilibrium concentration of ionized donors.
$D_n \equiv kT\mu_n/e$, diffusion constant for electrons.
$D_p \equiv kT\mu_p/e$, diffusion constant for holes.
$e \equiv$ magnitude of the electronic charge.
$E \equiv$ electrostatic field.
$E_a \equiv$ applied or asymptotic field.
$E_0 \equiv kT/eL_p$, characteristic field.
$\epsilon \equiv$ dielectric constant.
$f \equiv$ fraction of total current carried by holes.
$f_\epsilon \equiv$ fraction of total current carried by holes at an emitter.
$f^- \equiv$ fraction of total current carried by holes at a source, to the left.
$f^+ \equiv$ fraction of total current carried by holes at a source, to the right.
$F \equiv E/E_0$, reduced electrostatic field.
$g_0 \equiv$ thermal rate of generation of hole-electron pairs, per unit volume.
$G \equiv dP/dX$, reduced concentration gradient.
$G^0 \equiv$ value of G for $X = 0$.
$G^- \equiv$ limiting value of G at a source, approached from the left.
$G^+ \equiv$ limiting value of G at a source, approached from the right.
$H \equiv$ probability of thermal ionization of an unionized donor, per unit time.

I ≡ total current density.
I_n ≡ current density of electrons.
I_0 ≡ σE_0, characteristic current.
I_p ≡ current density of holes.
J ≡ $\frac{1}{e} I$, total carrier flow density.
J_n ≡ $-\frac{1}{e} I_n$, electron flow density.
J_p ≡ $\frac{1}{e} I_p$, hole flow density.
k ≡ Boltzmann's constant.
K ≡ probability per unit time of electron capture by an ionized donor, per unit electron concentration.
L_d ≡ $(kT\epsilon/8\pi n_i e^2)^{\frac{1}{2}}$, characteristic length associated with space charge in the steady state.
L_p ≡ $(kT\mu\tau/e)^{\frac{1}{2}}$, diffusion length for holes for time τ.
M ≡ $1 + \frac{b+1}{b} P$.
M_0 ≡ $1 + \frac{b+1}{b} P_0$.
μ ≡ μ_p ≡ mobility for holes.
μ_n ≡ mobility for electrons.
n ≡ concentration of electrons.
n_i ≡ thermal-equilibrium concentration of electrons (or holes) in the intrinsic semiconductor.
n_0 ≡ thermal-equilibrium concentration of electrons.
n_s ≡ saturation concentration excess of electrons, corresponding to complete ionization of donors.
N ≡ $n/(n_0 - p_0)$, reduced electron concentration for an n-type semiconductor.
p ≡ concentration of holes.
p_0 ≡ thermal-equilibrium concentration of holes.
P ≡ $p/(n_0 - p_0)$, reduced hole concentration for an n-type semiconductor.
ΔP ≡ $(p - p_0)/(n_0 - p_0)$, reduced concentration of added holes.
P_0 ≡ $p_0/(n_0 - p_0)$, reduced hole concentration at thermal equilibrium.
P^0 ≡ value of P for $X = 0$.
Q ≡ τ/τ_p, lifetime ratio.
R ≡ general recombination function, equal to $1 + aP/(1 + P_0)$ for mass-action and constant-mean-lifetime mechanisms combined.
ρ ≡ volume resistivity in ohm cm.
s ≡ differential transport velocity.
S ≡ $s/(D_p/\tau)^{\frac{1}{2}}$, reduced differential transport velocity.

σ $\equiv$ conductivity of semiconductor.
σ_0 $\equiv$ normal conductivity of semiconductor, with no added carriers.
Σ $\equiv \sigma/\sigma_0 = M/M_0$, reduced conductivity of semiconductor.
t $\equiv$ time variable.
T $\equiv$ temperature in degrees absolute.
τ $\equiv$ mean lifetime for holes for small added concentrations, in an n-type or in an intrinsic semiconductor.
τ_n $\equiv$ mean lifetime for electrons (concentration-dependent).
τ_p $\equiv$ mean lifetime for holes (concentration-dependent).
τ_r $\equiv$ mean lifetime for holes, for small added concentrations in an n-type semiconductor, due to mass-action recombination alone.
U $\equiv t/\tau \equiv$ reduced time variable.
W $\equiv eV/kT$, reduced electrostatic potential.
x $\equiv$ distance variable.
X $\equiv x/L_p$, reduced distance variable.
V $\equiv$ electrostatic potential.

Traveling-Wave Tubes

By J. R. PIERCE

[FOURTH INSTALLMENT]

CHAPTER XII

POWER OUTPUT

A THEORETICAL EVALUATION of the power output of a traveling-wave tube requires a theory of the non-linear behavior of the tube. In this book we have dealt with a linearized theory only. No attempt will be made to develop a non-linear theory. Some results of non-linear theory will be quoted, and some conclusions drawn from experimental work will be presented.

One thing appears clear both from theory and from experiment: the gain parameter C is very important in determining efficiency. This is perhaps demonstrated most clearly in some unpublished work of A. T. Nordsieck.

Nordsieck assumed:

(1) The same a-c field acts on all electrons.
(2) The only fields present are those associated with the circuit ("neglect of space charge").
(3) Field components of harmonic frequency are neglected.
(4) Backward-traveling energy in the circuit is neglected.
(5) A lossless circuit is assumed.
(6) C is small (it always is).

Nordsieck obtained numerical solutions for such cases for several electron velocities. He found the maximum efficiency to be proportonal to C by a factor we may call k. Thus, the power output P is

$$P = kCI_0V_0 \tag{12.1}$$

In Fig. 12.1, the factor k is plotted vs. the velocity parameter b. For an electron velocity equal to that of the unperturbed wave the fractional efficiency obtained is $3C$; for a faster electron velocity the efficiency rises to $7C$. For instance, if $C = .025$, $3C$ is 7.5% and $7C$ is 15%. For 1,600 volts 15 ma this means 1.8 or 3.6 watts. If, however, $C = 0.1$, which is attainable, the indicated efficiency is 30% to 70%.

Experimental efficiencies often fall very far below such figures, although some efficiencies which have been attained lie in this range. There are three apparent reasons for these lower efficiencies. First, small non-uniformities in wave propagation set up new wave components which abstract energy from the increasing wave, and which may subtract from the normal output. Second, when the a-c field varies across the electron flow, not all electrons

are acted on equally favorably. Third, most tubes have a central lossy section followed by a relatively short output section. Such tubes may overload so severely in the lossy section that a high level in the output section is never attained. There is not enough length of loss-free circuit to provide sufficient gain in the output circuit so that the signal can build up to maximum amplitude from a low level increasing wave. Other tubes with distributed loss suffer because the loss cuts down the efficiency.

Some power-series non-linear calculations made by L. R. Walker show that for fast velocities of injection the first non-linear effect should be an expansion, not a compression. Nordsieck's numerical solutions agree with this. A power series approach is inadequate in dealing with truly large-signal be-

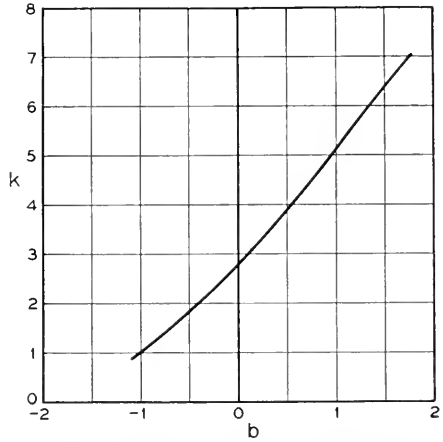

Fig. 12.1—The calculated efficiency is expressed as kC, where k is a function of the velocity parameter b. This curve shows k as given by Nordsieck's high-level calculations.

havior. In fact, Nordsieck's work shows that the power-series attack, if based on an assumption that there is no overtaking of electrons by electrons emitted later, must fail at levels much below the maximum output.

Further work by Nordsieck indicates that the output may be appreciably reduced by variation of the a-c field across the beam.

It is unfortunate that Nordsieck's calculations do not cover a wider range of conditions. Fortunately, unlikely as it might seem, the linear theory can tell us a little about what limitation of power we might expect. For instance, from (7.15) we have

$$\frac{v}{u_0} = -j\frac{\eta V}{u_0^2 \delta C}$$

$$\frac{v}{u_0} = -j\left(\frac{V}{2V_0}\right)\left(\frac{1}{\delta C}\right) \tag{12.2}$$

while from (7.16) we have

$$\frac{i}{I_0} = -\left(\frac{V}{2V_0}\right)\left(\frac{1}{\delta C}\right)^2 \tag{12.3}$$

We expect non-linear effects to become important when an a-c quantity is no longer small compared with a d-c quantity. We see that because $(1/\delta C)$ is large, $|\, i/I_0 \,|$ will be larger than $|\, v/u_0 \,|$.

The important non-linearity is a sort of over-bunching or limit to bunching. For instance, suppose we were successful in bunching the electron flow into very short pulses of electrons, as shown in Fig. 12.2 As the pulses approach zero length, the ratio of the peak value of the fundamental component of convection current to the average or d-c current I_0 approaches 2. We may, then, get some hint as to the variation of power output as various parameters are varied by letting $|\, i \,| = 2I_0$ and finding the variation of power in the circuit for an a-c convection current as we vary various parameters.

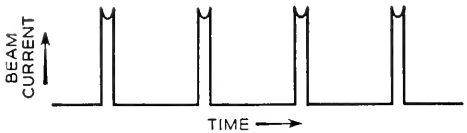

Fig. 12.2—If the electron beam were bunched into pulses short compared with a cycle, the peak value of the component of fundamental frequency would be twice the d-c current I_0.

Deductions made in this way cannot be more than educated guesses, but in the absence of non-linear calculations they are all we have.

From (7.1) we have for the *circuit field* associated with the *active mode* (neglecting the field due to space charge)

$$E = \frac{\Gamma^2 \Gamma_1 (E^2/\beta^2 P)}{2(\Gamma_1^2 - \Gamma^2)} \tag{12.4}$$

This relation is, of course, valid only for an electron convection current i which varies with distance as $\exp(-\Gamma z)$. For the power to be large for a given magnitude of current, E should be large. For a given value of i, E will be large if Γ is very nearly equal to Γ_1. This is natural. If Γ were equal to Γ_1, the natural propagation constant of the circuit, the contribution to the field by the current i in every elementary distance would have such phase as to add in phase with every other contribution.

Actually, Γ_1 and Γ cannot be quite equal. We have from (7.10) and (7.11)

$$-\Gamma_1 = \beta_e(-j - jCb - Cd) \tag{12.5}$$

$$-\Gamma = \beta_e(-j + jCy_1 + Cx_1) \tag{12.6}$$

For a physical circuit the attenuation parameter d must be positive while, for an increasing wave, x must be positive. We see that we may expect E to be greatest for a given current when d and x are small, and when y is nearly equal to the velocity parameter b.

Suppose we use (12.4) in expressing the power

$$P = \frac{E^2}{\beta^2(E^2/\beta^2 P)} = \left| \frac{\Gamma^4 \Gamma_1^2 (E^2/\beta^2 P)}{4\beta^2(\Gamma_1^2 - \Gamma^2)^2} \, i^2 \right|. \tag{12.7}$$

Here we identify β with $-j\Gamma_1$. Further, we use (2.43), (12.5) and (12.6), and assuming C to be small, neglect terms involving C compared with unity. We will further let i have a value

$$i = 2I_0 \tag{12.8}$$

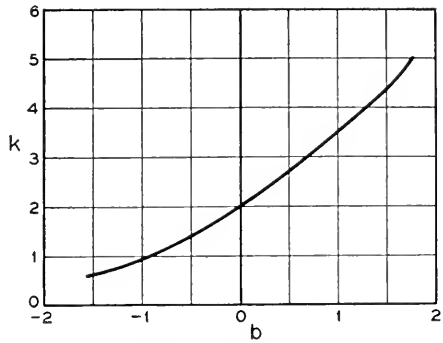

Fig. 12.3—An efficiency parameter k calculated by taking the power as that given by near theory for an r-f beam current with a peak value twice the d-c beam current.

We obtain

$$P = kCI_0V_0 \tag{12.9}$$

$$k = \frac{2}{(b + y)^2 + (x + d)^2} \tag{12.10}$$

We will now investigate several cases. Let us consider first the case of a lossless circuit ($d = 0$) and no space charge ($QC = 0$) and plot the efficiency factor k vs. b. The values of x and y are those of Fig. 8.1. Such a plot is shown in Fig. 12.3.

If we compare the curve of Fig. 12.3 with the correct curve of Nordsieck, we see that there is a striking qualitative agreement and, indeed, fair quantitative agreement. We might have expected on the one hand that the electron stream would never become completely bunched ($i = 2I_0$) and that, as it approached complete bunching, behavior would already be non-linear. This would tend to make (12.10) optimistic. On the other hand, even after i

attains its maximum value and starts to fall, power can still be transferred to the circuit, though the increase of field with distance will no longer be exponential. This makes it possible that the value of k given by (12.10) will be exceeded. Actually, the true k calculated by Nordsieck is a little higher than that given by (12.10).

Let us now consider the effect of loss. Figure 12.4 shows k from (12.10) vs. d for $b = QC = 0$. We see that, as might be expected, the efficiency falls as the loss is increased. C. C. Cutler has shown experimentally through unpublished work that the power actually falls off much more rapidly with d.

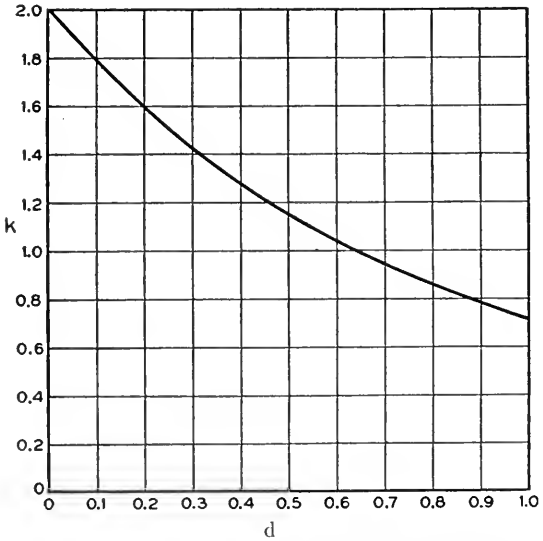

Fig. 12.4—The efficiency parameter k calculated as in Fig. 12.3 but for $b = 0$ (an electron velocity equal to the circuit phase velocity) and for various values of the attenuation parameter d. Experimentally, the efficiency falls off more rapidly as d is increased.

Finally, Fig. 12.5 shows k from (12.10) vs. QC, with $d = 0$ and b chosen to make x_1 a maximum. We see that there is a pronounced rise in efficiency as the space-charge parameter QC is increased.

J. C. Slater has suggested in Microwave Electronics a way of looking at energy production essentially based on observing the motions of electrons while traveling along with the speed of the wave. He suggests that the electrons might eventually be trapped and oscillate in the troughs of the sinusoidal field. If so, and if they initially have an average velocity Δv greater than that of the wave, they cannot emerge with a velocity lower than the velocity of the wave less Δv. Such considerations are complicated by the fact that the phase velocity of the wave in the large-signal region will not

be the same as its phase velocity in the small-signal region. It is interesting, however, to see what limiting efficiencies this leads to.

The initial electron velocity for the increasing wave is approximately

$$v_a = v_e(1 - y_1C) \tag{12.11}$$

where v_e is the phase velocity of the wave in the absence of electrons. The quantity y_1 is negative. According to Slater's reckoning, the final electron velocity cannot be less than

$$v_b = v_e(1 + y_1C) \tag{12.12}$$

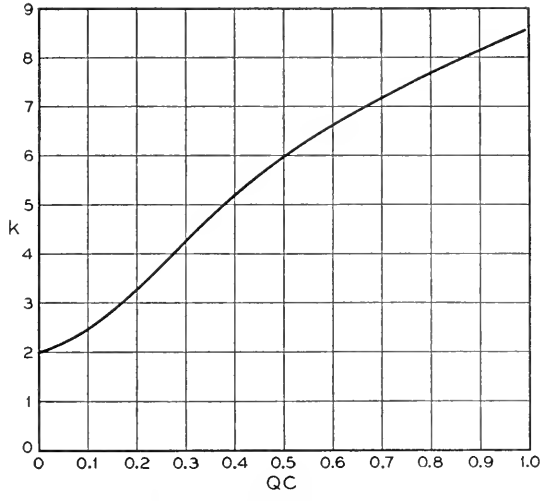

Fig. 12.5—The efficiency parameter k calculated as in Fig. 12.3, for zero loss and for an electron velocity which makes the gain of the increasing wave greatest, vs the space-charge parameter QC.

The limiting efficiency η accordingly will be, from considerations of kinetic energy

$$\eta = \frac{v_a^2 - v_b^2}{v_a^2}$$

$$\eta = \frac{4y_1C}{(1 - y_1C)^2}.$$

If $y_1C \ll 1$, very nearly

$$\eta = 4\,y_1C \tag{12.13}$$

We see that this also indicates an efficiency proportional to C. In Fig. 12.6 $4y_1$ is plotted vs. b for $QC = d = 0$. We see that this quantity ranges

from 2 for $b = 0$ up to 5 for larger values of b. It is surprising how well this agrees with corresponding values of 3 and 7 from Nordsieck's work. Moreover (12.13) predicts an increase in efficiency with increasing QC.

Thus, we may expect the efficiency to vary with C from several points of view.

It is interesting to consider what happens if at a given frequency we change the current. By changing the current while holding the voltage constant we increase both the input power and the efficiency, for C varies as $I_0^{1/3}$. Thus, in changing the current alone we would expect the power to vary as the 4/3 power of I_0

$$P \approx I_0^{4/3} \tag{12.14}$$

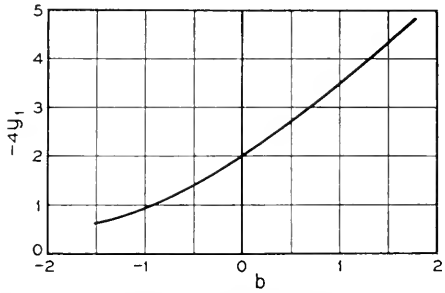

Fig. 12.6—According to a suggestion made by Slater, the velocity by which the electrons are slowed down cannot be greater than twice the difference between the electron velocity and the wave velocity. If we use the velocity difference given by the linear theory, for zero loss ($d = 0$) this would make the efficiency parameter k equal to $-4y_1$. Here $-4y_1$ is plotted vs b for $QC = 0$.

Here space charge has been neglected, and actually power may increase more rapidly with current than (12.14) indicates.

A variety of other cases can be considered. At a given voltage and current, C and the efficiency rise as the helix diameter is made smaller. However, as the helix diameter is made smaller it may be necessary to decrease the current, and the optimum gain will come at higher frequencies. For a given beam diameter, the magnetic focusing field required to overcome space-charge repulsion is constant if $I_0/V_0^{1/2}$ is held constant, and hence we might consider increasing the current as the 1/2 power of the voltage, and thus increasing the power input as the 3/2 power of the voltage. On the other hand, the magnetic focusing field required to correct initial angular deflections of electrons increases as the voltage is raised.

There is no theoretical reason why electrons should strike the circuit. Thus, it is theoretically possible to use a very high beam power in connection with a very fragile helix. Practically, an appreciable fraction of the beam current is intercepted by the helix, and this seems unavoidable for wave

lengths around a centimeter or shorter, for accurate focusing becomes more difficult as tubes are made physically smaller. Thus, in getting very high powers at ordinary wavelengths or even moderate powers at shorter wavelengths, filter type circuits which provide heat dissipation by thermal conduction may be necessary. We have seen that the impedance of such circuits is lower than that of a helix for the broadband condition (group velocity equal to phase velocity). However, high impedances and hence large values of C can be attained at the expense of bandwidth by lowering the group velocity. This tends to raise the efficiency, as do the high currents which are allowable because of good heat dissipation. However, lowering group velocity increases attenuation, and this will tend to reduce efficiency somewhat.

It has been suggested that the power can be increased by reducing the phase velocity of the circuit near the output end of the tube, so that the electrons which have lost energy do not fall behind the waves. This is a complicated but attractive possibility. It has also been suggested that the electrode which collects electrons be operated at a voltage lower than that of the helix.

The general picture of what governs and limits power output is fairly clear as long as C is very small. If attenuation near the output of the tube is kept small, and the circuit is constructed so as to approximate the requirement that nearly the same field acts on all electrons, efficiencies as large as 40% are indicated within the limitations of the present theory. With larger values of C it is not clear what the power limitation will be.

The usual traveling-wave tube would seem to have a serious competitor for power applications in the traveling-wave magnetron amplifier, which is discussed briefly in a later chapter.

CHAPTER XIII

TRANSVERSE MOTION OF ELECTRONS

Synopsis of Chapter

SO FAR WE HAVE taken into account only longitudinal motions of electrons. This is sufficient if the transverse fields are small compared to the longitudinal fields (as, near the axis of an axially symmetrical circuit) or, if a strong magnetic focusing field is used, so that transverse motions are inhibited. It is possible, however, to obtain traveling-wave gain in a tube in which the longitudinal field is zero at the mean position of the electron beam. For a slow wave, the electric field is purely transverse only along a plane. The transverse field in this plane forces electrons away from the plane and preferentially throws them into regions of retarding field, where they give up energy to the circuit. This mechanism is not dissimilar to that in the longitudinal field case, in which the electrons are moved longitudinally from their unperturbed positions, preferentially into regions of more retarding field.

Whatever may be said about tubes utilizing transverse fields, it is certainly true that they have been less worked on than longitudinal-field tubes. In view of this, we shall present only a simple analysis of their operation along the lines of Chapter II. In this analysis we take cognizance of the fact that the charge induced in the circuit by a narrow stream of electrons is a function not only of the charge per unit length of the beam, but of the distance between the beam and the circuit as well.

The factor of proportionality between distance and induced charge can be related to the field produced by the circuit. Thus, if the variation of V in the x, y plane (normal to the direction of propagation) is expressed by a function Φ, as in (13.3), the effective charge ρ_E is expressed by (13.8) and, if y is the displacement of the beam normal to the z axis, by (13.9) where Φ' is the derivative of Φ with respect to y.

The equations of motion used must include displacements normal to the z direction; they are worked out including a constant longitudinal magnetic focusing field. Finally, a combined equation (13.23) is arrived at. This is rewritten in terms of dimensionless parameters, neglecting some small terms, as (13.26)

$$j\delta - b = \frac{1}{\delta^2} + \frac{\alpha^2}{(\delta^2 + f^2)}.$$

Here δ and b have their usual meanings; α is the ratio between the transverse and longitudinal field strengths, and f is proportional to the strength of the magnetic focusing field.

In case of a purely transverse field, a new gain parameter D is defined. D is the same as C except that the longitudinal a-c field is replaced by the transverse a-c field. In terms of D, b and δ are redefined by (13.36) and (13.37), and the final equation is (13.38). Figures 13.5–13.10 show how the x's and y's vary with b for various values of f (various magnetic fields) and Fig. 13.11 shows how x_1, which is proportional to the gain of the increasing wave in db per wavelength, decreases as magnetic field is increased. A numerical example shows that, assuming reasonable circuit impedance, a magnetic field which would provide a considerable focusing action would still allow a reasonable gain.

The curves of Figs. 13.6–13.10 resemble very much the curves of Figs. 8.7–8.9 of Chapter VIII, which show the effect of space charge in terms of the parameter QC. This is not unnatural; in one case space charge forces tend to return electrons which are accelerated longitudinally to their undisturbed positions. In the other case, magnetic forces tend to return electrons which are accelerated transversely to their undisturbed positions. In each case the circuit field acts on an electron stream which can itself sustain oscillations. In one case, the oscillations are of a plasma type, and the restoring force is caused by space charge of the bunched electron stream; in the other case the electrons can oscillate transversely in the magnetic field with cyclotron frequency.

Let us, for instance, compare (7.13), which applies to purely longitudinal displacements with space charge, with (13.38), which applies to purely transverse fields with a longitudinal magnetic field. For zero loss ($d = 0$), (7.13) becomes

$$1 = (j\delta - b)(\delta^2 + 4QC)$$

While

$$1 = (j\delta - b)(\delta^2 + f^2) \tag{13.38}$$

describes the transverse case. Thus, if we let

$$4QC = f^2$$

the equations are identical.

When there is both a longitudinal and a transverse electric field, the equation for δ is of the fifth degree. Thus, there are five forward waves. For an electron velocity equal to the circuit phase velocity ($b = 0$) and for no attenuation, the two new waves are unattenuated.

If there is no magnetic field, the presence of a transverse field component merely adds to the gain of the increasing wave. If a small magnetic field is

imposed in the presence of a transverse field component, this gain is somewhat reduced.

13.1 Circuit Equation

Consider a tubular electrode connected to ground through a wire, shown in Fig. 13.1. Suppose we bring a charge Q into the tube from ∞. A charge Q will flow to ground through the wire. This is the situation assumed in the analysis of Chapter II. In Fig. 2.3 it is assumed that all the lines of force from the charge in the electron beam terminate on the circuit, so that the whole charge may be considered as impressed on the circuit.

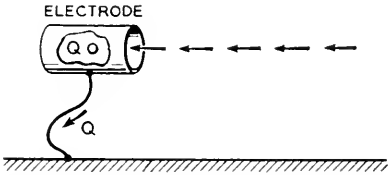

Fig. 13.1—When a charge Q approaches a grounded conductor from infinity and in the end all the lines of force from the charge end on the conductor, a charge Q flows in the grounding lead.

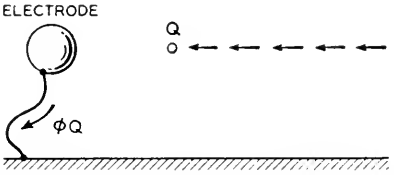

Fig. 13.2—If a charge Q approaches a conductor from infinity but in the end only part of the lines of force from the charge end on the conductor, a charge ΦQ flows in the grounding lead, where $\Phi < 1$.

Now consider another case, shown in Fig. 13.2, in which a charge Q is brought from ∞ to the vicinity of a grounded electrode. In this case, not all of the lines of force from the charge terminate on the electrode, and a charge ΦQ which is smaller than Q flows through the wire to ground.

We can represent the situation of Fig. 13.2 by the circuit shown in Fig. 13.3. Here C_2 is the capacitance between the charge and the electrode and C_1 is the capacitance between the charge and ground. We see that the charge ΦQ which flows to ground when a charge Q is brought to a is

$$\Phi Q = QC_2/(C_1 + C_2) \tag{13.1}$$

Now suppose we take the charge Q away and hold the electrode at a potential V with respect to ground, as shown in Fig. 13.4. What is the potential V_a at a? We see that it is

$$V_a = [C_2/(C_1 + C_2)]V = \Phi V \tag{13.2}$$

Thus, the same factor Φ relates the actual charge to the "effective charge" acting on the circuit and the actual circuit voltage to the voltage produced at the location of the charge.

We will not consider in this section the "space charge" voltage produced by the charge itself (the voltage at point a in Fig. 13.4).

The circuit voltage V we consider as varying as $\exp(-\Gamma z)$ in the direction of propagation. The voltage in the vicinity of the circuit is given by

$$V(x, y) = \Phi V \tag{13.3}$$

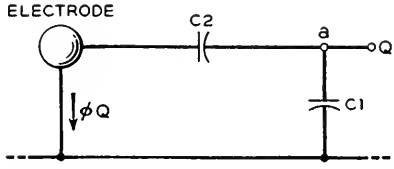

Fig. 13.3—The situation of Fig. 13.2 results in the same charge flow as if the charge were put on terminal a of the circuit shown, which consists of two capacitors of capacitances C_1 and C_2.

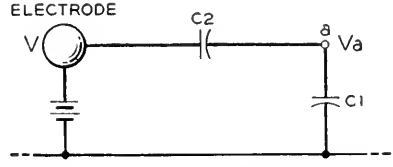

Fig. 13.4—A voltage V inserted in the ground lead divides across the condensers so that $V_a = \Phi V$, where Φ is the same factor which relates the charge flowing in the ground lead to the charge Q applied at a in Figs. 13.2 and 13.3.

Here x and y refer to coordinates normal to z and Φ is a function of x and y. We will choose x and y so

$$\partial\Phi/\partial x = 0 \tag{13.4}$$

Then

$$E_y = -V\partial\Phi/\partial y = -\Phi' V \tag{13.5}$$

$$\Phi' = \partial\Phi/\partial y \tag{13.6}$$

In (13.3), Φ will vary somewhat with Γ, but, as we are concerned with a small range only in Γ, we will consider Φ a function of y only.

From Chapter II we have

$$V = \frac{-\Gamma\Gamma_1 K i}{(\Gamma^2 - \Gamma_1^2)} \tag{2.10}$$

and

$$\rho = \frac{-j\Gamma i}{\omega}. \tag{2.18}$$

So that

$$V = \frac{-j\omega\Gamma_1 K\rho}{(\Gamma^2 - \Gamma_1^2)}. \tag{13.7}$$

In (13.7), it is assumed that $\Phi = 1$. If $\Phi \neq 1$, we should replace ρ in (13.7) by the a-c component of effective charge. The total effective charge ρ_E is

$$\rho_E = \Phi(\rho + \rho_0) \tag{13.8}$$

The term ρ_0 is included because Φ will vary if the y-position of the charge varies. To the first order, the a-c component ρ_E of the effective charge is,

$$\rho_E = \Phi\rho + \rho_0\Phi' y \tag{13.9}$$

$$\rho_E = \Phi\rho - (I_0/u_0)\Phi' y \tag{13.9}$$

Here y is the a-c variation in position along the y coordinate. Thus, if $\Phi \neq 0$, we have instead of (13.7)

$$V = \frac{-j\omega\Gamma_1 K(\Phi\rho - (I_0/u_0)\Phi' y)}{(\Gamma^2 - \Gamma_1^2)}. \tag{13.10}$$

This is the circuit equation we shall use.

13.2 Ballistic Equations

We will assume an unperturbed motion of velocity u_0 in the z direction, parallel to a uniform magnetic focusing field of strength B. As in Chapter II, products of a-c quantities will be neglected.

In the x direction, perpendicular to the y and z directions

$$d\dot{x}/dt = -\eta B\dot{y} \tag{13.11}$$

Assume that $\dot{x} = 0$ at $y = 0$. Then

$$\dot{x} = \eta B y \tag{13.12}$$

In the y direction we have

$$d\dot{y}/dt = \eta(B\dot{x} - E_y) \tag{13.13}$$

From (13.5) this is

$$d\dot{y}/dt = \eta(B\dot{x} + \Phi' V) \tag{13.14}$$

$$d\dot{y}/dt = \partial\dot{y}/\partial t + (\partial\dot{y}/\partial z)(dz/dt) \tag{13.15}$$

$$(d\dot{y}/dt) = u_0(j\beta_e - \Gamma)\dot{y} \tag{13.16}$$

We obtain from (13.16), (13.14) and (13.12)

$$(j\beta_e - \Gamma)y = -u_0\beta_m^2 y + \eta\Phi' V/u_0 \tag{13.17}$$

$$\beta_m = \eta B/u_0 \tag{13.18}$$

Here ηB is the cyclotron radian frequency and β_m is a corresponding propagation constant.

Now

$$\dot{y} = \partial y/\partial t - (\partial y/\partial z)(\partial z/\partial t) \tag{13.19}$$

$$\dot{y} = u_0(j\beta_e - \Gamma)y \tag{13.20}$$

From (13.20) and (13.17) we obtain

$$y = \frac{\Phi' V}{2\Gamma_0[(j\beta_e - \Gamma)^2 + \beta_m^2]}. \tag{13.21}$$

It is easily shown that the equation for ρ can be obtained exactly as in Chapter II. From (2.22) and (2.18) we have

$$\rho = \frac{I_0 \Gamma^2 \Phi V}{2u_0 V_0 (j\beta_e - \Gamma)^2}. \tag{13.22}$$

13.3 Combined Equation

From the circuit equation (13.10) and the ballistical equations (13.21) and (13.22) we obtain

$$1 = \frac{-j\beta_e \Gamma_1 \Gamma^2 \Phi^2 K I_0}{2V_0(\Gamma^2 - \Gamma_1^2)} \left[\frac{1}{(j\beta_e - \Gamma)^2} - \frac{(\Phi'/\Phi)^2}{\Gamma^2[(j\beta_e - \Gamma)^2 + \beta_m^2]} \right]. \tag{13.23}$$

The voltage at the beam is Φ times the circuit voltage, so the effective impedance of the circuit at the beam is Φ^2 times the circuit impedance. Thus

$$C^3 = \Phi^2 K I_0 / 4V_0 \tag{13.24}$$

It will be convenient to define a dimensionless parameter f specifying β_m and hence the magnetic field

$$f = \beta_m/\beta_e C \tag{13.25}$$

We will also use δ and b as defined earlier

$$-\Gamma = -j\beta_e + \beta_e C\delta$$

$$-\Gamma_1 = -j\beta_e - j\beta_e Cb$$

After the usual approximations, (13.23) yields

$$j\delta - b = \frac{1}{\delta^2} + \frac{\alpha^2}{(\delta^2 + f^2)} \tag{13.26}$$

$$\alpha^2 = (\Phi'/\beta_e \Phi)^2 \tag{13.27}$$

It is interesting to consider the quantity $(\Phi'/\beta_e\Phi)^2$ for typical fields. For

instance, in the two-dimensional electrostatic field in which the potential V is given by

$$V = Ae^{-\beta_e y}e^{-j\beta_e z} \tag{13.28}$$

$$\partial V/\partial y = -\beta_e V \tag{13.29}$$

and everywhere

$$\alpha^2 = (\Phi'/\beta_e\Phi)^2 = 1. \tag{13.30}$$

Relation (13.30) is approximately true far from the axis in an axially symmetrical field.

Consider a potential giving a purely transverse field at $y = 0$

$$V = Ae^{-j\beta_e z} \sinh \beta_e y \tag{13.31}$$

$$\frac{\partial V}{\partial y} = \beta A e^{-j\beta_e z} \cosh \beta_e y. \tag{13.32}$$

In this case, at $y = 0$

$$\alpha^2 = (\Phi'/\beta_e\Phi)^2 = \infty \tag{13.33}$$

In the case of a purely transverse field we let

$$D^3 = \frac{I_0\Phi'^2K}{4V_0\beta_e^2} \tag{13.34}$$

$$D^3 = (E_y^2/\beta^2P)(I_0/8V_0) \tag{13.35}$$

In (13.35), E_y is the magnitude of the y component of field for a power flow P, and β is the phase constant.

We then redefine δ and b in terms of D rather than C

$$-\Gamma = -j\beta_e + \beta_e D\delta \tag{13.36}$$

$$-\Gamma_1 = -j\beta_e - j\beta_e Db \tag{13.37}$$

and our equation for a purely transverse field becomes

$$1 = (j\delta - b)(\delta^2 + f^2) \tag{13.38}$$

In (13.38), δ and b are of course not the same as in (13.26) but are defined by (13.36) and (13.37).

13.4 Purely Transverse Fields

The case of purely transverse fields is of interest chiefly because, as was mentioned in Chapter X, it has been suggested that such tubes should have low noise.

In terms of x and y as usually defined

$$\delta = x + jy$$

equation (13.38) becomes

$$x[(x^2 - y^2 + f^2) - 2y(y + b)] = 0 \tag{13.39}$$

$$(y + b)(x^2 - y^2 + f^2) + 2x^2y + 1 = 0 \tag{13.40}$$

From the $x = 0$ solution of (13.39) we obtain

$$x = 0 \tag{13.41}$$

$$b = \frac{1}{y^2 - f^2} - y. \tag{13.42}$$

It is found that this solution obtains for large and small values of b. For very large and very small values of b, either

$$y \doteq -b \tag{13.43}$$

or

$$y \doteq \pm f \tag{13.44}$$

The wave given by (13.43) is a circuit wave; that given by (13.44) represents electrons traveling down the tube and oscillating with the cyclotron frequency in the magnetic field.

In an intermediate range of b, we have from (13.39)

$$x = \pm\sqrt{2y(y + b) - (f^2 - y^2)} \tag{13.45}$$

and

$$b = -2y \pm \sqrt{f^2 - 1/2y}. \tag{13.46}$$

For a given value of f^2 we can assume values of y and obtain values of b. Then, x can be obtained from (13.45). In Figs. 13.5–13.10, x and y are plotted vs. b for $f^2 = 0$, .5, 1, 4 and 10. It should be noted that x_1, the parameter expressing the rate of increase of the increasing wave, has a maximum at larger values of b as f is increased (as the magnetic focusing field is increased). Thus, for higher magnetic focusing fields the electrons must be shot into the circuit faster to get optimum results than for low fields. In Fig. 13.11, the maximum positive value of x is plotted vs. f. The plot serves to illustrate the effect on gain of increasing the magnetic field.

Let us consider an example. Suppose

$$\lambda = 7.5 \text{ cm}$$

$$D = .03$$

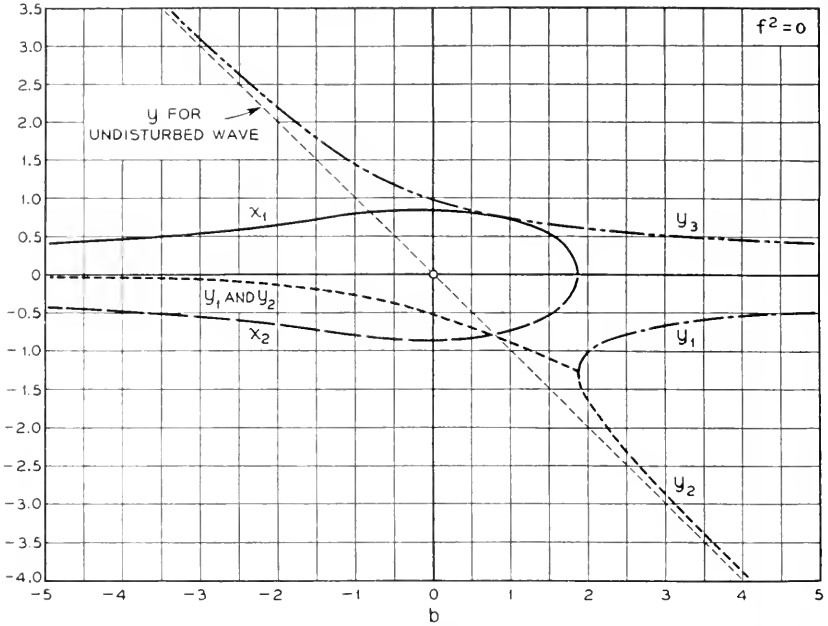

Fig. 13.5—The x's and y's for the three forward waves when the circuit field is purely transverse at the thin electron stream, for zero magnetic focusing field ($f^2 = 0$).

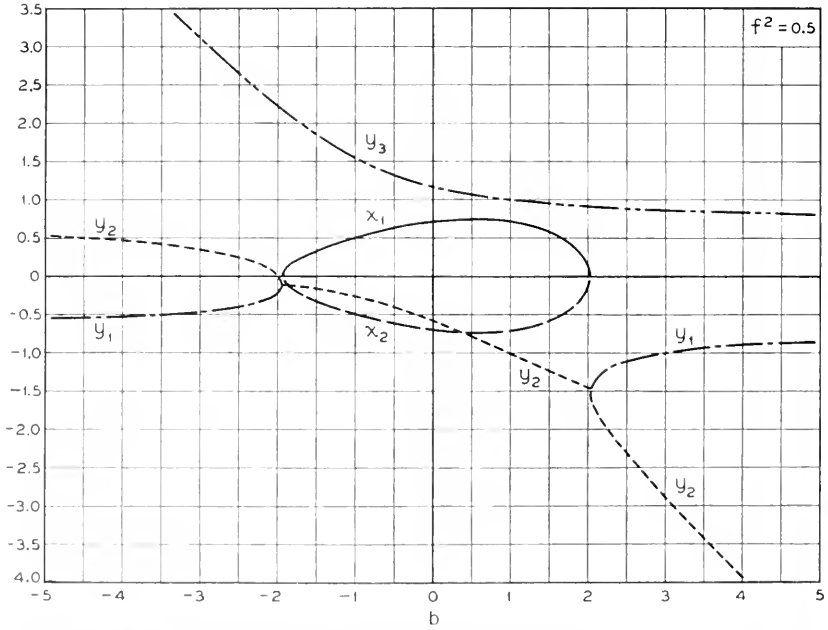

Fig. 13.6—Curves similar to those of Fig. 13.5 for a parameter $f^2 = 1$. The parameter f is proportional to the strength of the magnetic focusing field.

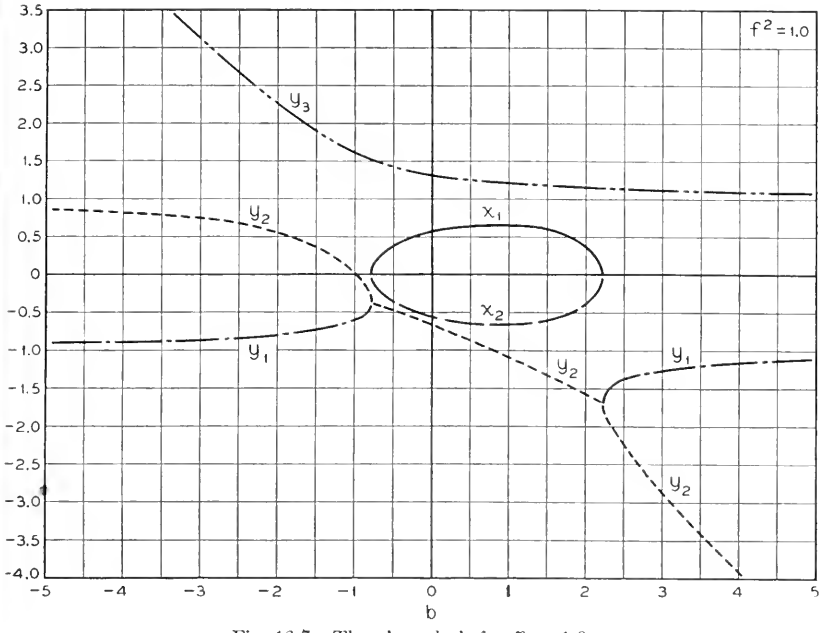

Fig. 13.7—The x's and y's for $f^2 = 1.0$.

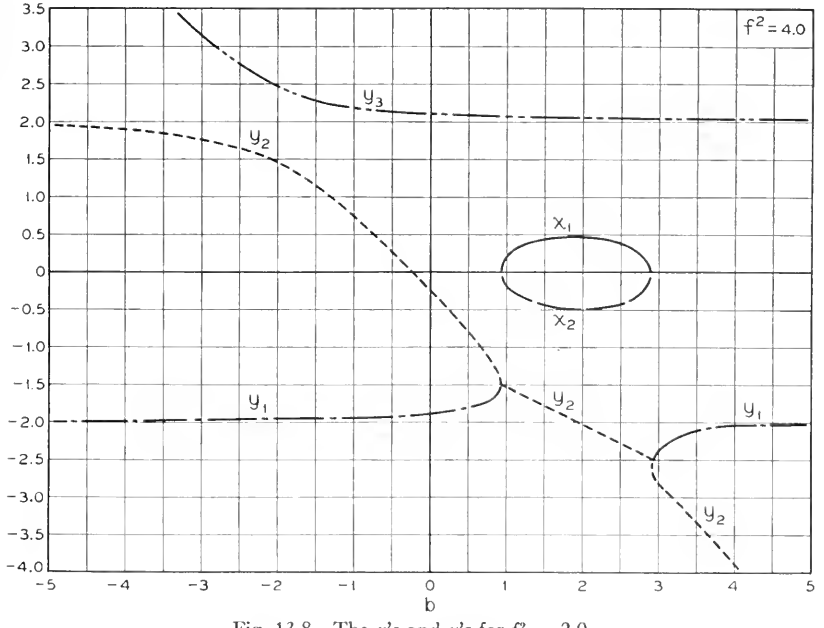

Fig. 13.8—The x's and y's for $f^2 = 2.0$.

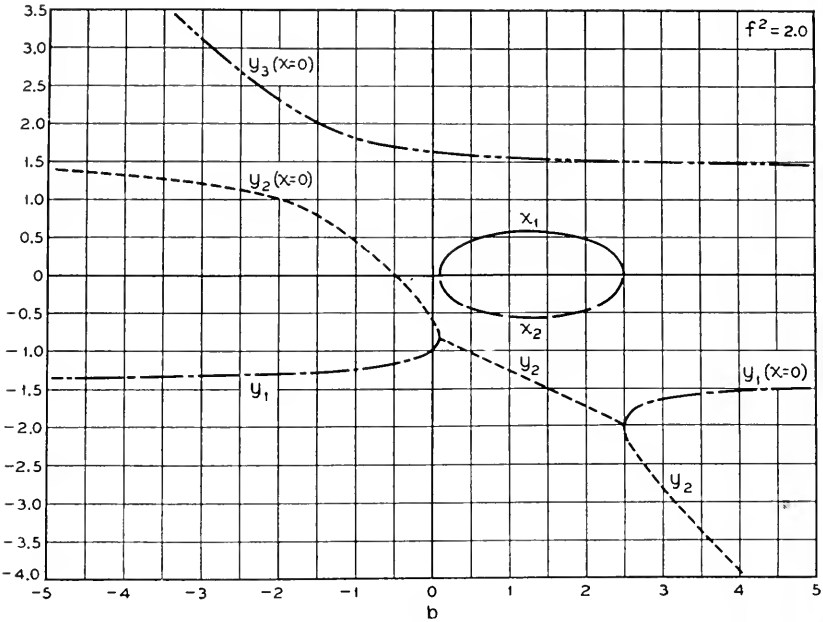

Fig. 13.9—The x's and y's for $f = 4.0$.

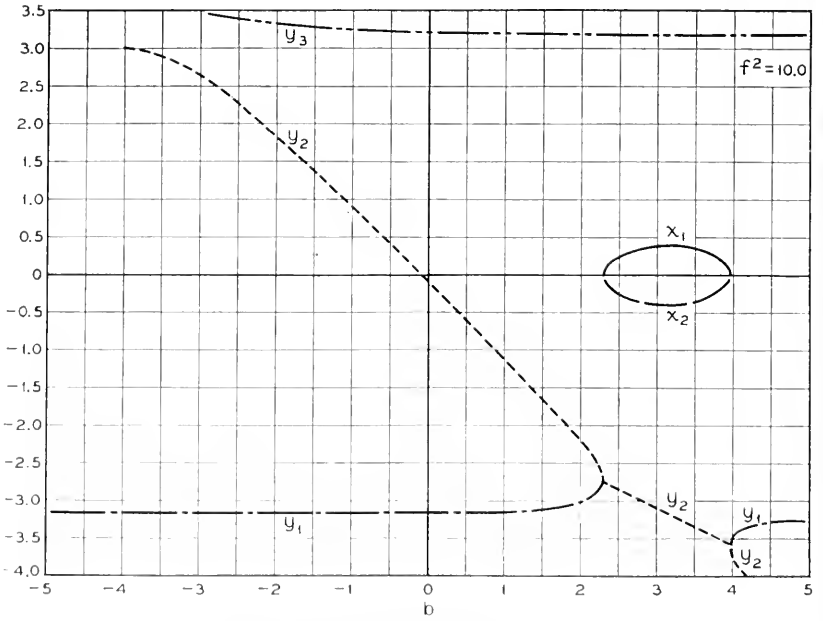

Fig. 13.10—The x's and y's for $f^2 = 10.0$.

These values are chosen because there is a longitudinal field tube which operates at 7.5 cm with a value of C (which corresponds to D) of about .03. The table below shows the ratio of the maximum value of x_1 to the maximum value of x_1 for no magnetic focusing field.

Magnestic Field in Gauss	f	x_1/x_{10}
0	0	1
50	1.17	.71
100	2.34	.50

A field of 50 to 100 gauss should be sufficient to give useful focusing action. Thus, it may be desirable to use magnetic focusing fields in transverse-

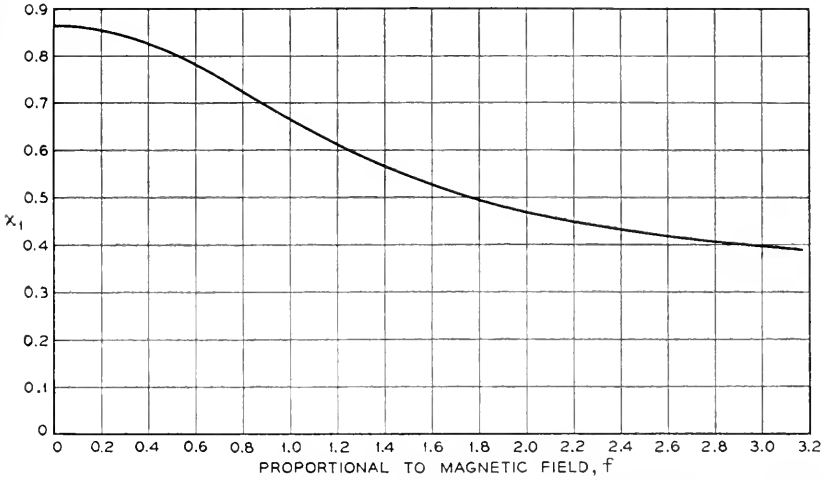

Fig. 13.11—Here x_1, the x for the increasing wave, is plotted vs f, which is proportional to the strength of the focusing field. The velocity parameter b has been chosen to maximize x_1. The ordinate x_1 is proportional to gain per wavelength.

field traveling-wave tubes. This will be more especially true in low-voltage tubes, for which D may be expected to be higher than .03.

13.5 Mixed Fields

In tubes designed for use with longitudinal fields, the transverse fields far off the axis approach in strength the longitudinal fields. The same is true of transverse field tubes far off the axis. Thus, it is of interest to consider equation (13.26) for cases in which α is neither very small nor very large, but rather is of the order of unity.

If the magnetic field is very intense so that f^2 is large, then the term containing α^2, which represents the effect of transverse fields, will be very small and the tube will behave much as if the transverse fields were absent.

Consideration of both terms presents considerable difficulty as (13.26) leads to five waves (5 values of δ) instead of three. The writer has attacked the problem only for the special case of $b = 0$. In this case we obtain from (13.26)

$$\delta = -j\left[\frac{1}{\delta^2} + \frac{\alpha^2}{\delta^2 + f^2}\right] \tag{13.47}$$

MacColl has shown[1] that the two "new" waves (waves introduced when $\alpha = 0$) are unattenuated and thus unimportant and uninteresting (unless, as an off-chance, they have some drastic effect in fitting the boundary conditions).

Proceeding from this information, we will find the change in δ as f^2 is increased from zero. From (13.47) we obtain

$$d\delta = j\left[\frac{2d\delta}{\delta^3} + \frac{2\alpha^2\,\delta d\delta}{(\delta^2 + f^2)^2} + \frac{\alpha^2\,df^2}{(\delta^2 + f^2)^2}\right] \tag{13.48}$$

Now, if $f = 0$

$$\delta^3 = -j(1 + \alpha^2) \tag{13.49}$$

If we use this in connection with (13.48) we obtain

$$d\delta = -\frac{\alpha^2}{3\delta}\,df^2 \tag{13.50}$$

For an increasing wave

$$\delta_1 = (1 + (\Phi'/\beta_e\Phi)^2)(\sqrt{3}/2 - j/2) \tag{13.51}$$

Hence, for the increasing wave

$$d\delta_1 = \frac{\alpha^2(-\sqrt{3}/2 - j/2)}{3(1 + \alpha^2)}\,df^2 \tag{13.52}$$

This shows that applying a small magnetic field tends to decrease the gain. This does not mean, however, that the gain with a longitudinal and transverse field and a magnetic field is less than the gain with the longitudinal field alone. To see this we assume that not f^2 but $(\Phi'/\beta_e\Phi)^2$ is small. Differentiating, we obtain

$$d\delta = -j\left[-\frac{2d\delta}{\delta^3} - \frac{2\alpha^2\,\delta d\delta}{(\delta^2 + f^2)^2} + \frac{d\alpha^2}{\delta^2 + f^2}\right] \tag{13.53}$$

If $\alpha = 0$

$$\delta^3 = -j \tag{13.54}$$

[1] J. R. Pierce, "Transverse Fields in Traveling-Wave Tubes," *Bell System Technical Journal*, Vol. 27, pp. 732–746.

and we obtain

$$d\delta = \frac{1}{3}\frac{\delta^3}{(\delta^2 + f^2)}\,d\alpha^2 \tag{13.55}$$

$$d\delta = \frac{-j}{3(\delta^2 + f^2)}\,d\alpha^2 \tag{13.56}$$

If we have a very large magnetic field ($f^2 \gg |\,\delta^2\,|$), then

$$d\delta = \frac{-j}{3f^2}\,d\alpha^2 \tag{13.57}$$

and the change in δ is purely reactive. If $f = 0$ (no magnetic field), from (13.55)

$$d\delta = \frac{\delta}{3}\,d\alpha^2 \tag{13.58}$$

Adding a transverse field component increases the magnitude of δ without changing the phase angle.

CHAPTER XIV

FIELD SOLUTIONS

Synopsis of Chapter

SO FAR, it has been assumed that the same a-c field acts on all electrons. This has been very useful in getting results, but we wonder if we are overlooking anything by this simplification.

The more complicated situation in which the variation of field over the electron stream is taken into account cannot be investigated with the same generality we have achieved in the case of "thin" electron streams. The chief importance we will attach to the work of this chapter is not that of producing numerical results useful in designing tubes. Rather, the chapter relates the appropriate field solutions to those we have been using and exhibits and evaluates features of the "broad beam" case which are not found in the "thin beam" case.

To this end we shall examine with care the simplest system which can reasonably be expected to exhibit new features. The writer believes that this will show qualitatively the general features of most or all "broad beam" cases.

The case is that of an electron stream of constant current density completely filling the opening of a double finned circuit structure, as shown in Fig. 14.1. The susceptance looking into the slots between the fins is a function of frequency only and not of propagation constant. Thus, at a given frequency, we can merely replace the slotted circuit members by susceptance sheets relating the magnetic field to the electric field, as shown in Fig. 14.2. The analysis is carried out with this susceptance as a parameter. Only the mode of propagation with a symmetrical field pattern is considered.

First, the case for zero current density is considered. The natural mode of propagation will have a phase constant β such that H_x/E_z for the central region is the same as H_x/E_z for the finned circuit. The solid curve of Fig. 14.3 shows a quantity proportional to H_x/E_z for the central space vs $\theta = \beta d$ (d defined by Fig. 14.1), a quantity proportional to β. The dashed line P represents H_x/E_z for a given finned structure. The intersections specify values of θ for the natural active modes of propagation to the left and to the right, and, hence, values of the natural phase constants.

The structure also has passive modes of propagation. If we assume fields which vary in the z direction as $\exp(\Phi/d)z$, H_x/E_z for the central

opening varies with Φ as shown in part in Fig. 14.4. A horizontal line representing a given susceptance of the finned structure will intersect the curve at an infinite number of points. Each intersection represents a passive mode which decays at a particular rate in the z direction and varies sinusoidally with a particular period in the y direction.

If the effect of the electrons in the central space is included, H_z/E_z for the central space no longer varies as shown in Fig. 14.3, but as shown in Fig. 14.5 instead. The curve goes off to $+\infty$ near a value of θ corresponding to a phase velocity near to the electron velocity. The nature of the modes depends on the susceptance of the finned structure. If this is represented by P_1, there are four unattenuated waves; for P_3 there are two unattenuated waves and an increasing and a decreasing wave. P_2 represents a transitional case.

Not the whole of the curve for the central space is shown on Fig. 14.5. In Fig. 14.6 we see on an expanded scale part of the region about $\theta = 1$, between the points where the curve goes through 0. The curve goes to $+\infty$ and repeatedly from $-\infty$ to $+\infty$, crossing the axis an infinite number of times as θ approaches unity. For any susceptance of the finned structure, this leads to an infinite number of unattenuated modes, which are space-charge waves; for these the amplitude varies sinusoidally with different periods across the beam. Not all of them have any physical meaning, for near $\theta = 1$ the period of cyclic variation across the beam will become small even compared to the space between electrons.

Returning to Fig. 14.1, we may consider a case in which the central space between the finned structures is very narrow (d very small). This will have the effect of pushing the solid curve of Fig. 14.5 up toward the horizontal axis, so that for a reasonable value of P (say, P_1, P_2 or P_3 of Fig. 14.5) there is no intersection. That is, the circuit does not propagate any unattenuated waves. In this case there are still an increasing and a decreasing wave. The behavior is like that of a multi-resonator klystron carried to the extreme of an infinite number of resonators. If we add resonator loss, the behavior of gain per wavelength with frequency near the resonant frequency of the slots is as shown in Fig. 14.7.

One purpose of this treatment of a broad electron stream is to compare its results with those of the previous chapters. There, the treatment considered two aspects separately: the circuit and the effect of the electrons.

Suppose that at $y = d$ in Fig. 14.1 we evaluate not H_x for the finned structure and for the central space separately, but, rather, the difference or discontinuity in H_x. This can be thought of as giving the driving current necessary to establish the field E_z with a specified phase constant. In Fig. 14.8, y_1 is proportional to this H_x or driving current divided by E_z. The dashed curve y_2 is the variation of driving current with θ or β which we have

used in earlier chapters, fitted to the true curve in slope and magnitude at $y = 0$. Over the range of θ of interest in connection with increasing waves, the fit is good.

The difference between H_x/E_z for the central space without electrons (Fig. 14.3) and H_x/E_z for the central space with electrons (Fig. 14.5) can be taken as representing the driving effect of the electrons. The solid curve of Fig. 14.9 is proportional to this difference, and hence represents the true effect of the electrons. The dashed curve is from the ballistical equation used in previous chapters. This has been fitted by adjusting the space-charge parameter Q only; the leading term is evaluated directly in terms of current density, beam width, β, and variation of field over the beam, which is assumed to be the same as in the absence of electrons.

Figure 14.10 shows a circuit curve (as, of Fig. 14.8) and an electronic curve (as, of Fig. 14.10). These curves contain the same information as the curves (including one of the dashed horizontal lines) of Fig. 14.5, but differently distributed. The intersections represent the modes of propagation.

If such curves were the approximate (dashed) curves of Figs. 14.8 and 14.9, the values of θ for the modes would be quite accurate for real intersections. It is not clear that "intersections" for complex values of θ would be accurately given unless they were for near misses of the curves. In addition, the complicated behavior near $\theta = 1$ (Fig. 14.6) is quite absent from the approximate electronic curve. Thus, the approximate electronic curve does not predict the multitude of unattenuated space-charge waves near $\theta = 1$. Further, the approximate expressions predict a lower limiting electron velocity below which there is no gain. This is not true for the exact equations when the electron flow fills the space between the finned structures completely.

It is of some interest to consider complex intersections in the case of near misses by using curves of simple form (parabolas), as in Fig. 14.11. Such an analysis shows that high gain is to be expected in the case of curves such as those of Fig. 14.10, for instance, when the circuit curve is not steep and when the curvature of the electronic curve is small. In terms of physical parameters, this means a high impedance circuit and a large current density.

14.1 The System and the Equations

The system examined is a two-dimensional one closely analogous to that of Fig. 4.4. It is shown in Fig. 14.1. It consists of a central space extending from $y = -d$ to $y = +d$, and arrays of thin fins separated by slots extending for a distance h beyond the central opening and short-circuited at the outer ends. An electron flow of current density J_0 amperes/m^2 fills the open space. It is assumed that the electrons are constrained by a strong magnetic field so that they can move in the z direction only.

We can simplify the picture a little. The open edges of the slots merely form impedance sheets.

From 4.12 we see that at $y = -d$

$$\frac{H_x}{E_z} = \frac{j\omega\epsilon}{\beta_0} \cot \beta_0 h \tag{14.1}$$

$$\frac{H_x}{E_z} = -jB \tag{14.2}$$

$$B = -\sqrt{\epsilon/\mu} \cot \beta_0 h \tag{14.3}$$

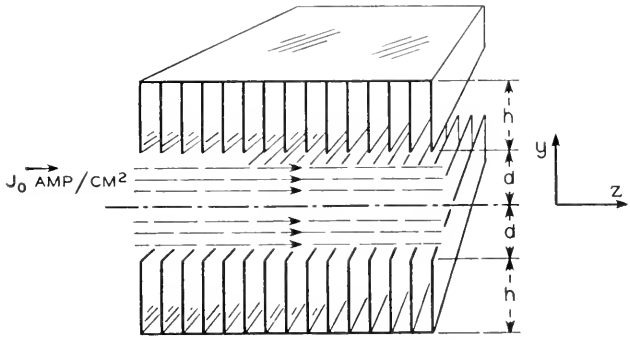

Fig. 14.1—Electron flow completely fills the open space between two finned structures. A strong axial magnetic field prevents transverse motions.

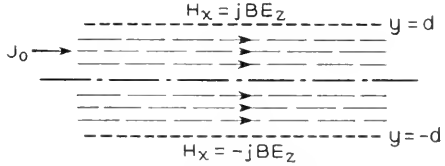

Fig. 14.2—In analyzing the structure of Fig. 14.1, the finned members are regarded as susceptance sheets.

for

$$\beta_0/\omega\epsilon = 1/c\epsilon = \sqrt{\mu/\epsilon} = 377 \text{ ohms} \tag{14.4}$$

Similarly, at $y = +d$,

$$\frac{H_x}{E_z} = jB \tag{14.5}$$

We can use B as a parameter rather than h. Thus, we obtain the picture of Fig. 14.2. This picture is really more general than Fig. 14.1, for it applies for any transverse-magnetic circuit outside of the beam.

Inside of the beam the effect of the electrons is to change the effective dielectric constant in the z direction. Thus, from (2.22) we have for the electron convection current

$$i = \frac{jJ_0\beta_e \Gamma V}{2V_0(j\beta_e - \Gamma)^2} \tag{2.22}$$

Now

$$E_z = -\frac{\partial V}{\partial z} = \Gamma V \tag{14.6}$$

so that

$$i = \frac{jJ_0\beta_e E_z}{2V_0(j\beta_e - \Gamma)^2} \tag{14.7}$$

The appearance of a voltage V in (2.22) and (14.6) does not mean that these relations are invalid for fast waves. In (2.22) the only meaning which need be given to V is that defined by (14.6), as it is the electric field as specified by (14.6) that was assumed to act on the electrons in deriving (2.22).

Let us say that the total a-c current density in the z direction, J_z, is

$$J_z = j\omega\epsilon_1 E_z \tag{14.8}$$

This current consists of a displacement current $j\omega\epsilon E_z$ and the current i, so that

$$J_z = j\omega\epsilon_1 E_z = j\omega\epsilon E_z\left(1 + \frac{J_0\beta_e}{2\epsilon\omega V_0(j\beta_e - \Gamma)^2}\right) \tag{14.9}$$

Hence

$$\epsilon_1/\epsilon = \left(1 + \frac{J_0\beta_e}{2\epsilon\omega V_0(j\beta_e - \Gamma)^2}\right) \tag{14.10}$$

This gives the ratio of the effective dielectric constant in the z direction to the actual dielectric constant. We will proceed to put this in a form which in the long run will prove more convenient.

Let us define a quanity β

$$\Gamma = j\beta \tag{14.11}$$

and a quantity A

$$A = \frac{J_0 d^2}{2\epsilon u_0 V_0} \tag{14.12}$$

And quantities θ and θ_e

$$\theta_e = \beta_e d = (\omega/u_0)d \tag{14.13}$$

$$\theta = \beta d \tag{14.14}$$

We recognize d as the half-width of the opening filled by electrons. Then

$$\epsilon_1/\epsilon = 1 - \frac{A}{(\theta_e - \theta)^2} \tag{14.15}$$

We can say something about the quantity A. From purely d-c considerations, the electron flow will cause a fall in d-c potential toward the center of the beam. Indeed, this is so severe for large currents that it sets a limit to the current density which can be transmitted. If we take V_0 and u_0 as values at $y = \pm d$ (the wall), the maximum value of A as defined by (14.12) is 2/3, and at this maximum value the potential at $y = 0$ is $V_0/4$. This is inconsistent with the analysis, in which V_0 and u_0 are assumed to be constant across the electron flow. Thus, for the current densities for which the analysis is valid, which are the current densities such as are usually used in traveling-wave tubes

$$A \ll 1 \tag{14.16}$$

In the a-c analysis we will deal here only with the symmetrical type of wave in which $E_z(+y) = E_z(-y)$. The work can easily be extended to cover cases for which $E_z(+y) = -E_z(-y)$. We assume

$$H_x = H_0 \sinh \gamma y e^{-j\beta z} \tag{14.17}$$

From Maxwell's equations

$$j\omega\epsilon E_y = \frac{\partial H_x}{\partial z} = -j\beta H_0(\sinh \gamma y)e^{-j\beta z}$$

$$E_y = -\frac{\beta}{\omega\epsilon} H_0(\sinh \lambda y)e^{-j\beta z} \tag{14.18}$$

Similarly

$$j\omega\epsilon_1 E_z = -\frac{\partial H_x}{\partial y} = -\gamma H_0(\cosh \gamma y)e^{-j\beta z}$$

$$E_z = \frac{j\gamma}{\omega\epsilon_1} H_0(\cosh \gamma y)e^{-j\beta z} \tag{14.19}$$

We must also have

$$-j\omega\mu H_x = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}$$

$$-j\omega\mu H_0 e^{-j\beta z} \sinh \gamma y = \frac{j\gamma^2}{\omega\epsilon_1} H_0 e^{-j\beta z} \cosh \gamma y - \frac{j\beta^2}{\omega\epsilon} H_0 e^{-j\beta z} \sinh \gamma y$$

$$\gamma^2 = (\epsilon_1/\epsilon)(\beta^2 - \beta_0^2) \tag{14.20}$$

$$\beta_0^2 = \omega^2\mu\epsilon = \omega^2/c^2 \tag{14.21}$$

Now, from (14.17), (14.19) and (14.20)

$$\frac{H_x}{E_z} = \frac{-j\omega\epsilon(\epsilon_1 - \epsilon)\tanh\,[(\epsilon_1 - \epsilon)^{1/2}(\beta^2 - \beta_0^2)^{1/2}\,y]}{(\epsilon_1/\epsilon)^{1/2}(\beta^2 - \beta_0^2)^{1/2}} \tag{14.22}$$

But

$$\omega\epsilon = (\omega/c)(c\epsilon) = \beta_0\sqrt{\epsilon/\mu} \tag{14.23}$$

Hence

$$\frac{H_x}{E_z} = \frac{-j\sqrt{\epsilon/\mu}(\epsilon_1 - \epsilon)^{1/2}\beta_0\tanh\,[(\epsilon_1 - \epsilon)^{1/2}(\beta^2 - \beta_0^2)^{1/2}y]}{(\beta^2 - \beta_0^2)^{1/2}} \tag{14.24}$$

At $y = d$, (14.5) must apply. From (14.24) we can write

$$P = -\frac{(\epsilon_1 - \epsilon)^{1/2}\tanh\,[(\epsilon_1 - \epsilon)^{1/2}(\theta^2 - \theta_0^2)^{1/2}]}{(\theta^2 - \theta_0^2)^{1/2}} \tag{14.25}$$

Here θ is given by (14.14)

$$\theta_0 = \beta_0 d = (\omega/c)d \tag{14.26}$$

and P is given by

$$P = B/\beta_0 d\sqrt{\epsilon/\mu} = B/\theta_0\sqrt{\epsilon/\mu} \tag{14.27}$$

Thus, θ_0 expresses d in radians at free-space wavelength and P is a measure of the wall reactance, the susceptance rising as B rises.

14.2 Waves in the Absence of Electrons

In this section we will consider (14.25) in the case in which there are no electrons and $\epsilon_1 - \epsilon = 1$. In this case (14.25) becomes

$$P = -\frac{\tanh\,(\theta^2 - \theta_0^2)^{1/2}}{(\theta^2 - \theta_0^2)^{1/2}} \tag{14.28}$$

Suppose we plot the right-hand side of (14.28) vs θ for real values of θ_1 corresponding to unattenuated waves. In Fig. 14.3 this has been done for $\theta_0 = 1.10$. For $\theta_0 > \pi/2$ the behavior near the origin is different, but in cases corresponding to actual traveling wave tubes $\theta_0 < \pi/2$.

Intersections between a horizontal line at height P and the curve give values of θ representing unattenuated waves. We see that for the case which we have considered, in which $\theta_0 < \pi/2$ and $\theta_0 \cot \theta_0 > 1$, there are unattenuated waves if

$$P > -\tan\theta_0\,\theta_0 \tag{14.29}$$

For $P = -\infty$ (no slot depth and no wall reactance) the system for $\theta_0 < \pi/2$ constitutes a wave guide operated below cutoff frequency for the type of

wave we have considered. If we increase P ($|P|$ decreasing; the inductive reactance of the walls increasing) this finally results in the propagation of a wave. There are two intersections, at $\theta = \pm\theta_1$, representing propagation to the right and propagation to the left. The variation of θ_1 with P is such that as P is increased (made less negative) θ_1 is increased; that is, the greater is P (the smaller $|P|$), the more slowly the wave travels.

There is another set of waves for which θ is imaginary; these represent passive modes which do not transmit energy but merely decay with distance. In investigating these modes we will let

$$\theta = j\Phi \tag{14.30}$$

so that the waves vary with z as

$$e^{(\Phi/d)z} \tag{14.31}$$

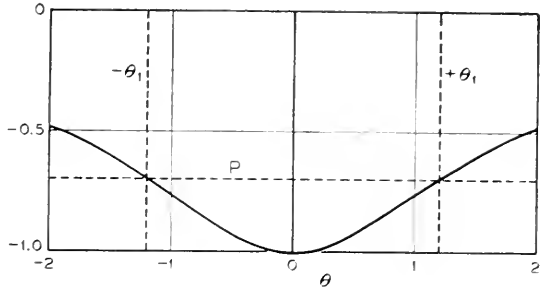

Fig. 14.3—The structure of Fig. 14.1 is first analyzed in the absence of an electron stream. Here a quantity proportional to H_z/E_z at the susceptance sheet is plotted vs $\theta = \beta d$, a quantity proportional to the phase constant β. The solid curve is for the inner open space; the dashed line is for the susceptance sheet. The two intersections at $\pm\theta_1$ correspond to transmission of a forward and a backward wave.

Now (14.28) becomes

$$P = -\tan(\Phi^2 + \theta_0^2)^{1/2}/(\Phi^2 + \theta_0^2)^{1/2} \tag{14.32}$$

In Fig. 14.4 the right-hand side of (14.28) has been plotted vs Φ, again for $\theta_0 = 1/10$.

Here there will be a number of intersections with any horizontal line representing a particular value of P (a particular value of wall susceptance), and these will occur at paired values of Φ which we shall call $\pm\Phi_n$. The corresponding waves vary with distance as $\exp(\pm\Phi_n z/d)$.

Suppose we increase P. As P passes the point $-(\tan\theta_0)\,\theta_0$, Φ^n for a pair of these passive waves goes to zero; then for P just greater than $-(\tan\theta_0)/\theta_0$ we have two active unattenuated waves, as may be seen by comparing Figs. 14.4 and 14.3.

14.3 Waves in the Presence of Electrons

In this section we deal with the equations

$$P = \frac{-(\epsilon_1/\epsilon)^{1/2} \tanh [(\epsilon_1/\epsilon)^{1/2}(\theta^2 - \theta_0^2)^{1/2}]}{(\theta^2 - \theta_0^2)^{1/2}} \tag{14.25}$$

and

$$\epsilon_1/\epsilon = 1 - \frac{A}{(\theta_e - \theta)^2} \tag{14.15}$$

We consider cases in which the electron velocity is much less than the velocity of light; hence

$$\theta_e \gg \theta_0 \tag{14.33}$$

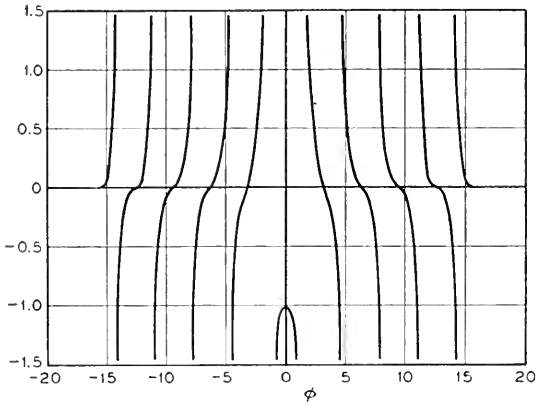

Fig. 14.4—If a quantity proportional to H_x/E_z at the edge of the central region is plotted vs $\Phi = -j\theta$, this curve is obtained. There are an infinite number of intersections with a horizontal line representing the susceptance of the finned structure. These correspond to passive modes, for which the field decays exponentially with distance away from the point of excitation.

In Fig. 14.5, the right-hand side of (14.25) has been plotted vs. θ for $\theta_e = 10\ \theta_0$, corresponding to an electron velocity 1/10 the speed of light. Values of $\theta = 1/10$ and $A = 1/100$ have been chosen merely for convenience.* The curve has not been shown in the region from $\theta = .9$ to $\theta = 1.1$, where ϵ_1/ϵ is negative, and this region will be discussed later.

For a larger value of P(| P | small), P_1 in Fig. 14.5, there are 4 intersections corresponding to 4 unattenuated waves. The two outer intersections obviously correspond to the "circuit" waves we would have in the absence of electrons. The other two intersections near $\theta = .9\theta_e$ and $\theta = 1.1\theta_e$ we call electronic or space-charge waves.

* At a beam voltage $V_0 = 1{,}000$ and for $d = 0.1$ cm, $A = 1/100$ means a current density of about 330 ma/cm^2, which is a current density in the range encountered in practice.

For instance, increasing P to values larger than P_1 changes θ for the circuit waves a great deal but scarcely alters the two "electronic wave" values of θ, near $\theta = \theta_e(1 \pm 0.1)$. On the other hand, for large values of P the values of θ for the electronic waves are approximately

$$\theta = \theta_e \pm \sqrt{A} \tag{14.34}$$

Thus, changing A alters these values, but changing A has little effect on the values of θ for the circuit waves.

Now, the larger the P the slower the circuit wave travels; and, hence, for large values of P the electrons travel faster than the circuit wave. Our narrow-beam analysis also indicated two circuit waves and two unattenuated electronic waves for cases in which the electron speed is much larger than the speed of the increasing wave. It also showed, however, that, as the difference between the electron speed and the speed of the unperturbed

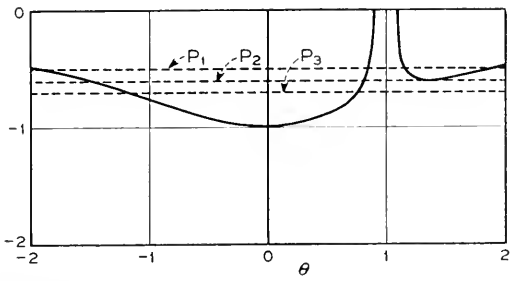

Fig. 14.5—When electrons are present in the open space of the circuit of Fig. 14.1, the curves of Fig. 14.3 are modified as shown here. The nature of the waves depends on the relative magnitude of the susceptance of the finned structure, which is represented by the dashed horizontal lines. For P_1, there are four unattenuated waves, for P_3, two unattenuated waves and an increasing wave and a decreasing wave. Line P_2 represents a transition between the two cases.

wave was made less, a pair of waves appeared, one increasing and one decreasing. This is also the case in the broad beam case.

In Fig. 14.5, when P is given the value indicated by P_2, an "electronic" wave and a "circuit" wave coalesce; this corresponds to y_1 and y_2 running together at $b = (3/2)(2)^{1/3}$ in Fig. 8.1. For a somewhat smaller value of P, such as P_3, there will be a pair of complex values of θ corresponding to an increasing wave and a decreasing wave. We may expect the rate of increase at first to rise and then to fall as P is gradually decreased from the value P_2, corresponding to the rise and fall of x_1 as b is decreased from $(3/2)(2)^{1/3}$ in Fig. 8.1.

It is interesting to know whether or not these increasing waves persist down to $P = -\infty$ (no inductance in the walls). When $P = -\infty$, the only way (14.25) can be satisfied is by

$$\coth\left((\epsilon_1/\epsilon)^{1/2}(\theta^2 - \theta_0^2)^{1/2}\right) = 0 \tag{14.35}$$

This will occur only if

$$\begin{aligned}(\epsilon_1/\epsilon)^{1/2}(\theta^2 - \theta_0^2)^{1/2} &= j\left(n\pi + \frac{\pi}{2}\right)\\ (\epsilon_1/\epsilon)(\theta^2 - \theta_0^2) &= -\left(n\pi + \frac{\pi}{2}\right)^2\end{aligned} \tag{14.36}$$

Let

$$\theta = u + jw \tag{14.37}$$

From (14.37), (14.36) and (14.15)

$$\left[1 - \frac{A}{((\theta_e - u) + jw)^2}\right]((u + jw)^2 - \theta_0^2) = -\left(n\pi + \frac{\pi}{2}\right)^2 \tag{14.38}$$

If we separate the real and imaginary parts, we obtain

$$\begin{aligned}&[(A - 1)(\theta_e - u)^2 - (A + 1)w^2](u^2 - w^2 - \theta_0^2)\\ &\qquad - 4Auw^2(\theta_e - u) = [(\theta_e - u)^2 + w^2]\left(n\pi + \frac{\pi}{2}\right)^2\end{aligned} \tag{14.39}$$

$$w(u[(\theta_e - u)^2 + w^2] - A[(\theta_e - u)^2 - w^2] + (\theta_e - u)(u^2 - w^2 - \theta_0^2)) = 0 \tag{14.40}$$

The right-hand side of (14.39) is always positive. Because always $A < 1$, the first term on the left of (14.39) is always negative if $u^2 > (w^2 + \theta_0^2)$, which will be true for slow rates of increase. Thus, for very small values of w, (14.39) cannot be satisfied. Thus, it seems that there are no waves such as we are looking for, that is, slow waves ($u \ll c$). It appears that the increasing waves must disappear or be greatly modified when P approaches $-\infty$.

So far we have considered only four of the waves which exist in the presence of electrons. A whole series of unattenuated electron waves exist in the range

$$\theta_e - \sqrt{A} < \theta < \theta_e + \sqrt{A}$$

In this range $(\epsilon_1/\epsilon)^{1/2}$ is imaginary, and it is convenient to rewrite (14.25) as

$$P = \frac{(-\epsilon_1/\epsilon)^{1/2} \tan [(-\epsilon_1/\epsilon)^{1/2}(\theta^2 - \theta_0^2)^{1/2}]}{(\theta^2 - \theta_0^2)^{1/2}} \tag{14.41}$$

The chief variation in this expression over the range considered is that due to variation in $(-\epsilon_1/\epsilon)^{1/2}$. For all practical purposes we may write

$$P = \frac{(-\epsilon_1/\epsilon)^{1/2} \tan [(-\epsilon_1/\epsilon)^{1/2}(\theta_e^2 - \theta_0)^{1/2}]}{(\theta_e^2 - \theta_0^2)^{1/2}} \tag{14.42}$$

Near $\theta = \theta_e$, the tangent varies with infinite rapidity, making an infinite number of crossings of the axis.

In Fig. 14.6, the right-hand side of (14.41) has been plotted for a part of the range $\theta = 0.90\,\theta_e$ to $\theta = 1.10\,\theta_e$. The waves corresponding to the intersections of the rapidly fluctuating curve with a horizontal line representing P are unattenuated space-charge waves. The nearer θ is to θ_e, the larger $(-\epsilon_1/\epsilon)$ is. The amplitude of the electric field varies with y as

$$\cosh\,(j(-\epsilon_1/\epsilon)^{1/2}(\beta^2 - \beta_0^2)^{1/2}y) = \cos\,((-\epsilon_1/\epsilon)^{1/2}(\beta^2 - \beta_0^2)^{1/2}y) \quad (14.45)$$

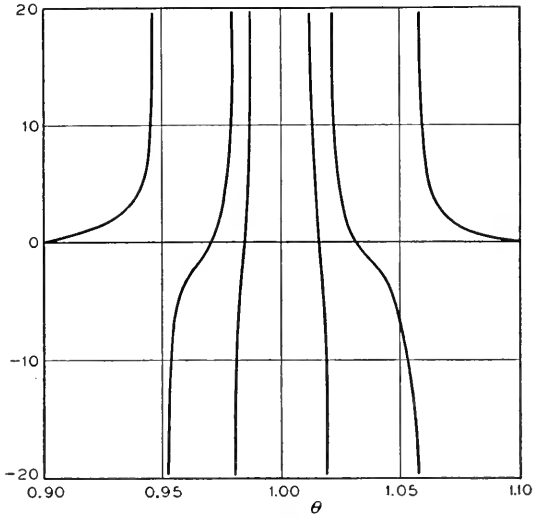

Fig. 14.6—The curve for the central region is not shown completely in Fig. 14.5. A part of the detail around $\theta = 1$, which means a phase velocity equal to the electron velocity, is shown in Fig. 14.6. The curve crosses the axis, and any other horizontal line, an infinite number of times (only some of the branches are shown). Thus, there is a large number of unattenuated "space charge" waves. For these, the amplitude varies sinusoidally in the y direction. Some of these have no physical reality, because the wavelength in the y direction is short compared with the space between electrons.

For small values of $|\theta - \theta_e|$ the field fluctuates very rapidly in the y direction, passing through many cycles between $y = 0$ and $y = d$. For very small values of $|\theta - \theta_e|$ the solution does not correspond to any actual physical problem: spreads in velocity in any electron stream, and ultimately the discrete nature of electron flow, preclude the variations indicated by (14.45).

The writer cannot state definitely that there are not increasing waves for which the real part of θ lies between $\theta_e - \sqrt{A}$ and $\theta_e + \sqrt{A}$, but he sees no reason to believe that there are.

There are, however, other waves which exhibit both attenuation and

propagation. The roots of (14.32) are modified by the introduction of the electrons. To show this effect, let Φ_n be a solution of (14.32), and $j(\Phi_n + \delta)$ be a solution of (14.25). The waves considered will thus vary with distance as

$$e^{[(\Phi_n+\delta)/d]z} \tag{14.43}$$

We see that we must have

$$\begin{aligned}(\epsilon_1/\epsilon)^{1/2}\,(\Phi_n^2 + \theta_0^2)^{1/2}\,\cot\,(\Phi_n^2 + \theta_0^2)^{1/2} \\ = ((\Phi_n + \delta)^2 + \theta_0^2)^{1/2}\,\cot\,[(\epsilon_1/\epsilon)^{1/2}((\Phi_n + \delta)^2 + \theta_0^2)^{1/2}]\end{aligned} \tag{14.44}$$

$$(\epsilon_1/\epsilon)^{1/2} = \left(1 - \frac{A}{(\theta_e - j\Phi_n + \delta)^2}\right)^{1/2} \tag{14.15a}$$

As $A \ll 1$, it seems safe to neglect δ in (14.15a) and to expand, writing

$$(\epsilon_1/\epsilon)^{1/2} = 1 - \alpha \tag{14.46}$$

$$\alpha = \frac{A}{2(\theta_e - j\Phi_n)^2} = \frac{A[(\theta_e^2 - \Phi_n^2) + 2j\theta_e\Phi_n]}{2(\theta_e^2 + \Phi_n^2)^2} \tag{14.47}$$

If $|\,\delta\,| \ll \Phi_n$, we may also write

$$((\Phi_n + \delta)^2 + \theta_0^2)^{1/2} = \frac{\Phi_n\,\delta}{(\Phi_n^2 + \theta_0^2)^{1/2}} + (\Phi_n^2 + \theta_0^2)^{1/2} \tag{14.48}$$

We thus obtain, if we neglect products of δ and α

$$\begin{aligned}(1 - \alpha)\,\cot\,(\Phi_n^2 + \theta_0^2)^{1/2} = \left[1 + \frac{\Phi_n\,\delta}{(\Phi_n^2 + \theta_0^2)^{1/2}}\right]\cot\,(\Phi_n^2 + \theta_0^2)^{1/2} \\ - \left(\frac{\Phi_n\,\delta}{(\Phi_n^2 + \theta_0^2)^{1/2}} - \alpha\right)\csc^2\,(\Phi_n^2 + \theta_0^2)^{1/2}\end{aligned} \tag{14.49}$$

Solving this for δ, we obtain

$$\delta = -\frac{(\Phi_n^2 + \theta_0^2)^{1/2}}{\Phi_n}\left[\frac{\cos\,(\Phi_n^2 + \theta_0^2)^{1/2} + \csc\,(\Phi_n^2 + \theta_0^2)^{1/2}}{\cos\,(\Phi_n^2 + \theta_0^2)^{1/2} - \csc\,(\Phi_n^2 + \theta_0^2)^{1/2}}\right]\alpha \tag{14.50}$$

$$\begin{aligned}\delta = \left[\frac{(\theta_e^2 - \Phi_n^2)}{\Phi_n(\theta_e^2 + \Phi_n^2)^2} + j\,\frac{2\theta_e}{(\theta_e^2 + \Phi_n^2)^2}\right] \\ \cdot \left[\frac{\csc^2\,(\Phi_n^2 + \theta_0^2)^{1/2} + \cos\,(\Phi_n^2 + \theta_0^2)^{1/2}}{\csc\,(\Phi_n^2 + \theta_0^2)^{1/2} - \cos\,(\Phi_n^2 + \theta_0^2)^{1/2}}\right]\frac{A(\theta_0^2 + \Phi_n^2)^{1/2}}{2}\end{aligned} \tag{14.51}$$

As the waves vary with distance as $\exp\,[(\pm\,\Phi_n + \delta)z/d]$, this means that all modified waves travel in the $-z$ direction, and very fast, for the imaginary part of δ, which is inversely proportional to the phase velocity, will be small.

These backward-traveling waves cannot give gain in the $+z$ direction, and could give gain in the $-z$ direction only under conditions similar to those discussed in Chapter XI.

14.4 A Special Type of Solution

Consider (14.25) in a case in which

$$\theta_0 \ll \theta_e \tag{14.52}$$

$$\theta_e \ll 1 \tag{14.53}$$

In this case in the range

$$\theta < \theta_e - \sqrt{A} \quad \text{and} \quad \theta > \theta_e + \sqrt{A} \tag{14.54}$$

we can replace the hyperbolic tangent by its argument, giving

$$P = -(\epsilon_1/\epsilon) = \frac{A}{(\theta_e - \theta)^2} - 1. \tag{14.55}$$

This can be solved for θ, giving

$$\theta = \theta_e \mp \sqrt{A/(P+1)} \tag{14.56}$$

If

$$P < -1$$

Then θ will be complex and there will be a pair of waves, one increasing and one decreasing. We note that, under these circumstances, there is no circuit wave, either with or without electrons.

What we have is in essence an electron stream passing through a series of inductively detuned resonators, as in a multi-resonator klystron. Thus, the structure is in essence a distributed multi-resonator klystron, with lossless resonators. If the resonators have loss, we can let

$$P = (-jG + B)/\theta_0\sqrt{\epsilon/\mu} \tag{14.57}$$

where G is the resonant conductance of the slots. In this case, (14.56) becomes

$$\theta = \theta_e \pm \left(\frac{A\theta\sqrt{\epsilon/\mu}}{-jG + (B + \theta_0\sqrt{\epsilon/\mu})}\right)^{1/2} \tag{14.58}$$

Near resonance we can assume G is a constant and that B varies linearly with frequency. Accordingly, we can show the form of the gain of the increasing wave by plotting vs. frequency the quantity g

$$g = \text{Im}(-j + \omega/\omega_0)^{-1/2} \tag{14.59}$$

In Fig. 14.7, g is plotted vs. ω/ω_0.

14.5 Comparison with Previous Theory

We will compare our field solution with the theory presented earlier by comparing separately circuit effects and electronic effects.

14.6a *Comparison of Circuit Equations*

According to Chapter VI the field induced in an active mode by the current i should be

$$E_z = \frac{\Gamma^2 \Gamma_1 (E^2/\beta^2 P)}{2(\Gamma_1^2 - \Gamma^2)}\, i$$

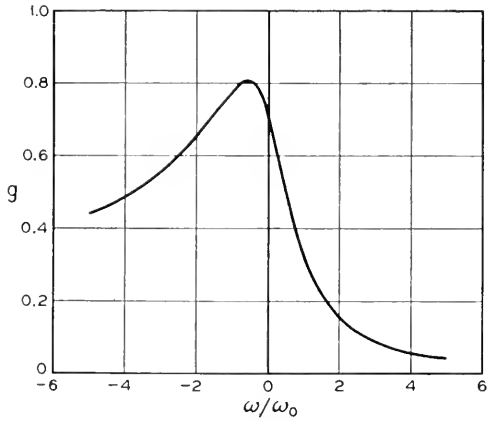

Fig. 14.7—In a plot such as that of Fig. 14.5, the horizontal line for the fins may not intersect the solid line for the central space at all. Particularly, this will be true as the central space is made very narrow. There will still be an increasing and a decreasing wave, however. Suppose, now, that the finned structure is lossy. We find that the gain in db of the increasing wave will vary with frequency as shown. Here ω_0 is the resonant frequency of the slots in the finned structure.

whence

$$E_z = \frac{j\theta^2 R}{(\theta_1^2 - \theta^2)}\, i \tag{14.60}$$

where R is a positive constant proportional to $(E^2/\beta^2 P)$.

Suppose that in Fig. 14.2 we have at $y = d$ not only the current jBE_z flowing in the wall admittance, but an additional current i given by (14.60) as well. Then instead of (14.28) we have

$$\frac{1}{\theta_0(\sqrt{\epsilon/\mu})}\,\frac{i}{jE_z} + P = -\frac{\tanh\,(\theta^2 - \theta_0^2)^{1/2}}{(\theta^2 - \theta_0^2)^{1/2}} \tag{14.61}$$

For simplicity, let $\theta_0 \ll \theta$. Then we obtain from (14.61)

$$i = j\theta_0 \sqrt{\epsilon/\mu} \left(-P - \frac{\tanh \theta}{\theta} \right) E_z \tag{14.62}$$

We must identify this with (14.60). Thus, over the range considered, we must have approximately

$$(\theta_1^2/\theta^2 - 1)/R = \theta_0\sqrt{\epsilon/\mu}(P + (\tanh\theta)/\theta) \tag{14.63}$$

At $\theta = \theta_1$, we must have both sides zero, so that

$$P = -(\tanh\theta_1)/\theta_1 \quad \text{and} \tag{14.64}$$

$$(1 - (\theta_1/\theta)^2)/R = \sqrt{\epsilon/\mu}((\tanh\theta_1)/\theta_1 - (\tanh\theta)/\theta) \tag{14.65}$$

Taking the derivative with respect to θ

$$\frac{2\theta_1^2}{\theta^3 R} = \theta_0\sqrt{\epsilon/\mu}\left(-\frac{\operatorname{sech}^2\theta}{\theta} + \frac{\tanh\theta}{\theta^2}\right) \tag{14.66}$$

These must be equal at $\theta = \theta_1$, so that

$$1/R = (1/2)(\theta_0\sqrt{\epsilon/\mu})\left(\frac{\tanh\theta_1}{\theta_1} - \operatorname{sech}^2\theta_1\right) \tag{14.67}$$

Thus, according to the methods of Chapter VI, our circuit equation should be

$$\left(\frac{1}{\theta_0\sqrt{\epsilon/\mu}}\right)\frac{i}{jE_z} = (1/2)\left(\frac{\tanh\theta_1}{\theta_1} - \operatorname{sech}^2\theta_1\right)(1 - (\theta_1/\theta)^2) \tag{14.68}$$

Using (14.64), the correct equation (14.62) becomes

$$\left(\frac{1}{\theta_0\sqrt{\epsilon/\mu}}\right)\frac{i}{jE_z} = \frac{\tanh\theta_1}{\theta_1} - \frac{\tanh\theta}{\theta} \tag{14.69}$$

In a typical traveling-wave tube, we might have

$$\theta_1 = 2.5$$

In Fig. 14.8, the right-hand side of (14.69) is plotted as a solid line and the right-hand side of (14.68) is plotted as a dashed line for $\theta_1 = 2.5$.

14.5b *Electronic Comparison*

Consider (14.25), which is the equation with electrons. For simplicity, let $\theta_0 \ll \theta$, so that

$$\frac{B}{\sqrt{\epsilon/\mu}\,\theta_0} = P = -\frac{(\epsilon_1/\epsilon)^2\tanh[(\epsilon_1/\epsilon)^{1/2}\theta]}{\theta} \tag{14.70}$$

For no electrons we would have

$$\frac{B}{\theta_0\sqrt{\epsilon/\mu}} = P = -\frac{\tanh\theta}{\theta} \tag{14.71}$$

Thus, if we wish we may write (14.70) in the form

$$P_e = -\frac{\tanh \theta}{\theta} - P \tag{14.72}$$

where

$$P_e = (1/\theta)[(\epsilon_1/\epsilon)^{1/2} \tanh [(\epsilon_1/\epsilon)^{1/2} \theta - \tanh \theta] \tag{14.73}$$

The quantities on the right of (14.72) refer to the circuit in the absence of electrons; if there are no electrons $P_e = 0$ and (14.72) yields the circuit

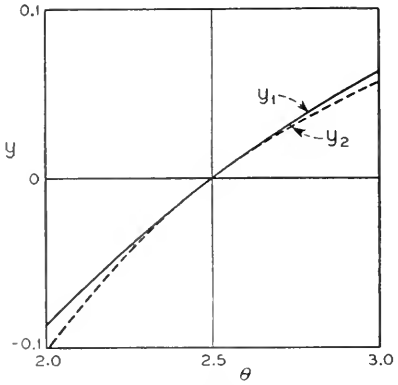

Fig. 14.8—Suppose we compare the circuit admittance for the structure of Fig. 14.1 with that used in earlier calculations. Here the solid curve is proportional to the difference of the H_x's for the finned structure and for the central space (the impressed current) divided by E_z. The dashed curve is the simple expression (6.1) used earlier fitted in magnitude and slope.

waves. Thus, P_e may be regarded as the equivalent of an added current i at the wall, such that

$$\frac{i}{jE_z} = \theta \sqrt{\epsilon/\mu}\, P_e \tag{14.74}$$

Now, the root giving the increasing wave, the one we are most interested in, occurs a little way from the pole, where $(\epsilon_1/\epsilon)^{1/2}$ may be reasonably large if θ is large. It would seem that one of the best comparisons which could be made would be that between the approximate analysis and a very broad beam case, for which θ is very large. In this case, we may take approximately, away from $\theta = \theta_e$

$$\tanh [(\epsilon_1/\epsilon)^{1/2} \theta] = \tanh \theta = 1 \tag{14.75}$$

$$P_e = (1/\theta)[(\epsilon_1/\epsilon)^{1/2} - 1$$

$$P_e = (1/\theta)\left[\left(1 - \frac{A}{(\theta_e - \theta)^2}\right)^{1/2} - 1\right] \tag{14.76}$$

Let us expand in terms of the quantity $A/(\theta_e - \theta)^2$, assuming this to be small compared with unity. We obtain

$$P_e = \frac{A}{2\theta(\theta_e - \theta)^2}\left[1 + \frac{A}{4(\theta_e - \theta)^2} + \cdots\right] \tag{14.77}$$

The theory of Chapter VII is developed by assuming that all electrons are acted on by the same a-c field. When this is not so, it is applied approximately by using an "effective current" or "effective field" as in Chapter IV; either of these concepts leads to the same averaging over the electron flow. An effective current can be obtained by averaging over the flow the current density times the square of the field, evaluated in the absence of electrons, and dividing by the square of the field at the reference position. This is equivalent to the method used in evaluating the effective field in Chapter III.

In the device of Fig. 14.2, if we take as a reference position $y = \pm d$, the effective current I_0 per unit depth

$$I_0 = \frac{J_0 \int_0^d \cosh^2(\gamma y)\, dy}{\cosh^2 \gamma d} \tag{14.78}$$

$$I_0 = (Jd/2)\left(\frac{\tanh \gamma d}{\gamma d} + \operatorname{sech}^2 \gamma d\right) \tag{14.79}$$

This is the effective current associated with the half of the flow from $y = 0$ to $y = d$. Here γ is the value for no electrons. For $\theta \ll \beta$, $\gamma = \beta$. For large values of θ, then

$$I_0 = J_0 d/2\theta \tag{14.80}$$

Now, the corresponding a-c convection current per unit depth will be:

$$i = -j\frac{I_0 \beta_e}{2V_0(\beta_e - \beta)^2} E \tag{14.81}$$

Here E is the total field acting on the electrons in the z-direction. From (7.1) we see that we assumed this to be the field due to the circuit (the first term in the brackets) plus a quantity which we can write

$$E_{z1} = \frac{j\beta^2}{\omega C_1} i \tag{14.82}$$

Accordingly

$$E = E_z + E_{z1} \tag{14.83}$$

and we can write i

$$i = -j \frac{I_0 \beta_e}{2V_0(\beta_e - \beta)^2} \left(E_z + \frac{j\beta^2}{\omega C_1} i \right) \tag{14.84}$$

$$i = \frac{jI_0 \theta_e \, dE_z}{2V_0[K - (\theta_e - \theta)^2]} \tag{14.85}$$

Here K is a parameter specifying the value of $\beta^2/\omega C_1$. As (14.85) need hold over only a rather small range of β, and C is not independent of β, we will regard K as a constant.

The parameter P_e corresponding to (14.85) is

$$P_e = \frac{I_0 \, d(\theta_e/\theta_0)}{2\sqrt{\epsilon/\mu}\, V_0} [K - (\theta_e - \theta)^2]^{-1} \tag{14.86}$$

Now, from (14.80), for large values of θ

$$\frac{I_0 \, d(\theta_e/\theta_0)}{2\sqrt{\epsilon/\mu}\, V_0} = \frac{J_0 \, d^2(\theta_e/\theta_0)}{4\sqrt{\epsilon/\mu}\, \theta V_0} \tag{14.87}$$

As

$$\sqrt{\epsilon/\mu} = \epsilon/\sqrt{\mu\epsilon} = \epsilon c,$$

$$\theta_e/\theta_0 = c/u_0,$$

and

$$A = \frac{J_0 \, d^2}{2\epsilon u_0 V_0} \tag{14.12}$$

$$P_e = \frac{A}{2\theta[K - (\theta_e - \theta)^2]} \tag{14.88}$$

Let us now expand (14.88) assuming K to be very small

$$P_e = \frac{A}{2\theta(\theta_e - \theta)^2} \left[1 + \frac{K}{(\theta_e - \theta)^2} + \cdots \right] \tag{14.89}$$

If we let

$$K = A/4 \tag{14.90}$$

we see that these first two terms agree with the expansion of the broad-beam expression, (14.77). The leading term was not adjusted; the space-charge parameter K was, since there is no other way of evaluating the parameter in this case.

In Fig. 14.9, the value of θP_e as obtained, actually, from (14.73) rather than (14.76), is plotted as a solid line and the value corresponding to the

earlier theory, from (14.86) with K adjusted according to (14.88), is plotted as a dashed line, for

$$A = 0.01$$

$$\theta_e = 8$$

We see that (14.88), which involves the approximations made in our earlier calculations concerning traveling-wave tubes, is a remarkably good fit to the broad-beam expression derived from field theory up very close to the points $(\theta_e - \theta) = A$, which are the boundaries between real and imaginary arguments of the hyperbolic tangent and correspond to the points where the ordinate is zero in Fig. 14.5.

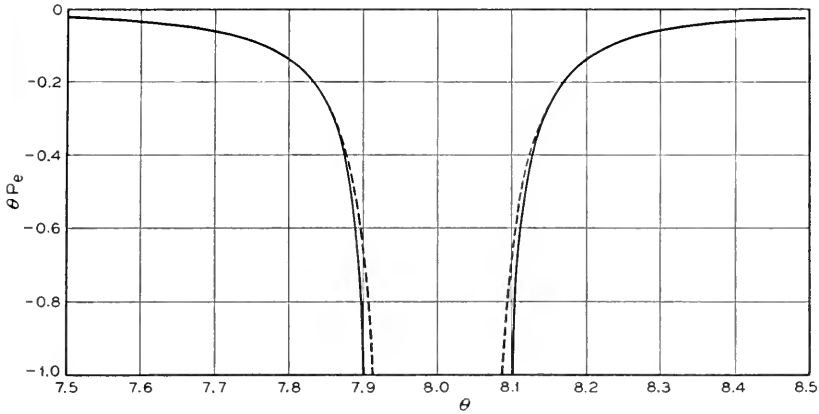

Fig. 14.9—These curves compare an exact electronic susceptance for the broad beam case (solid curve) with the approximate expression used earlier (dashed curve). In the approximate expression, the "effective current" was evaluated, not fitted; the space-charge parameter was chosen to give a fit.

Over the range in which the argument of the hyperbolic tangent in the correct expression is imaginary, the approximate expression of course exhibits none of the complex behavior characteristics of the correct expression and illustrated by Fig. 14.6. From (14.88) we see that the multiple excursions of the true curve from $-\infty$ to $+\infty$ are replaced in the approximate curve by a single dip down toward 0 and back up again. R. C. Fletcher has used a method similar to that explained above in computing the effective helix impedance and the effective space-charge parameter Q for a solid beam inside of a helically conducting sheet. His work, which is valuable in calculating the gain of traveling-wave tubes, is reproduced in Appendix VI.

14.5c *The Complex Roots*

The propagation constants represent intersections of a circuit curve such as that shown in Fig. 14.8 and an electronic curve such as that shown in Fig.

14.9. The propagation constants obtained in Chapters II and VIII represent such intersections of approximate circuit and electronic curves, such as the dotted lines of Fig. 14.8 and 14.9. Propagation constants obtained by field solutions represent intersections of the more nearly exact circuit and electronic curves such as the solid curves of Figs. 14.8 and 14.9.

If we plot a circuit curve giving

$$(1/\theta_0 \sqrt{\epsilon/\mu})(i/jE_z)$$

as given by (14.65) (the right-hand side of 14.75) and an electronic curve giving

$$(1/\theta_0 \sqrt{\epsilon/\mu})(i/jE_z) = P_e$$

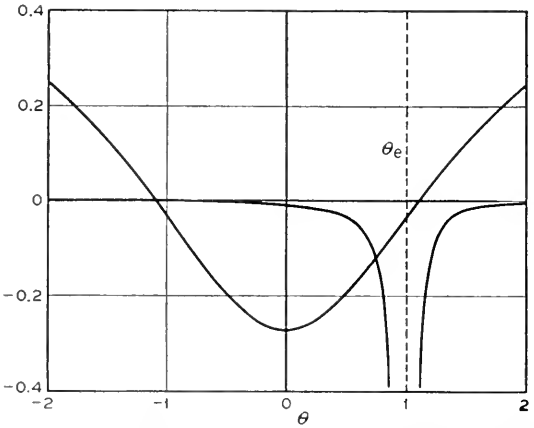

Fig. 14.10—The curves of Fig. 14.5 may be replaced by those of Fig. 14.6. Here the curve which is concave upward represents the circuit susceptance and the other curve represents the electronic susceptance (as in Fig. 14.9).

as given by (14.73) (the left-hand side of (14.72)), the plot, which is shown in Fig. 14.10, contains the same information as the plot of Fig. 14.5 for which θ_0, θ_e and A are the same. In Fig. 14.10, however, one curve represents the circuit without electrons and the other represents the added effect of the electrons.

We have seen that the approximate expressions of Chapter VII fit the broad-beam curves well for real propagation constants (real values of θ) (Fig. 14.8 and 14.9). Hence, we expect that complex roots corresponding to the increasing waves which are obtained using the approximate expressions will be quite accurate when the circuit curve is not too far from the electronic curve for real values of θ; that is, when the parameters (electron velocity, for instance) do not differ too much from those values for which the circuit curve is tangent to the electronic curve.

Unfortunately, the behavior of a function for values of the variables far

from those represented by its intersection with the real plane may be very sensitive to the shape of the intersection with the real plane. Thus, we would scarcely be justified by the good fit of the approximations represented in Figs. 14.8 and 14.9 in assuming that the complex roots obtained using the approximations will be good except when they correspond to a near approach of the electronic and circuit curves, as in Fig. 14.10.

In fact, using the approximate curves, we find that the increasing wave vanishes for electron velocities less than a certain lower limiting velocity. This corresponds to cutting by the circuit curve of the dip down from $+\infty$ of the approximate electronic curve (the dip is not shown in Fig. 14.9). This is not characteristic of the true solution. An analysis shows, however,

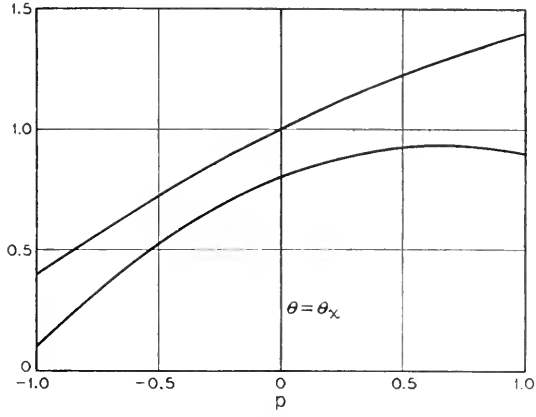

Fig. 14.11—Complex roots are obtained when curves such as those of Fig. 14.10 do not have the number of intersections required (by the degree of the equation) for real values of the abscissa and ordinate. In this figure, two parabolas narrowly miss intersect ng. Suppose these represent circuit and electronic susceptance curves. We find that the gain of the increasing wave will increase with the square root of the separation at the abscissa of equal slopes, and inversely as the square root of the difference in second derivatives.

that there will be a limiting electron velocity below which there is no increasing wave if there is a charge-free region between the electron flow and the circuit.

14.6 Some Remarks About Complex Roots

If we examine our generalized circuit expression (14.60) we see that the circuit impedance parameter $(E^2/\beta^2 P)$ is inversely proportional to the slope of the circuit curve at the point where it crosses the horizontal axis. Thus, low-impedance circuits cut the axis steeply and high-impedance circuits cut the axis at a small slope.

We cannot go directly from this information to an evaluation of gain in terms of impedance; the best course in this respect is to use the methods of

Chapter VIII. We can, however, show a relation between gain and the properties of the circuit and electronic curves for cases in which the curves almost touch (an electron velocity just a little lower than that for which gain appears). Suppose the curves nearly touch at $\theta = \theta_x$, as indicated in Fig. 14.11. Let

$$\theta = \theta_x + p \tag{14.91}$$

Let us represent the curves for small values of p by the first three terms of a Taylor's series. Let the ordinate y of the circuit curve be given by

$$y = a_1 + b_1 p + c_1 p^2 \tag{14.92}$$

and let the ordinate of the electronic curve be given by

$$y = a_2 + b_2 p + c_2 p^2 \tag{14.93}$$

Then, at the intersection

$$(c_1 - c_2)p^2 + (b_1 - b_2)p + (a_1 - a_2) = 0$$

$$p = -(1/2)\frac{b_1 - b_2}{c_1 - c_2} \pm j\sqrt{\frac{(a_1 - a_2)}{(c_1 - c_2)} - \frac{(b_1 - b_2)^2}{4(c_1 - c_2)^2}} \tag{14.94}$$

If we choose θ_x as the point at which the slopes are the same

$$b_1 - b_2 = 0 \tag{14.95}$$

$$p = \pm j\sqrt{\frac{(a_1 - a_2)}{(c_1 - c_2)}} \tag{14.96}$$

and we see that the imaginary part of p increases with the square root of the separation, and at a rate inversely proportional to the difference in second derivatives. This is exemplified by the behavior of x_1 and x_2 for b a little small than $(3/2)(2)^{1/3}$ in Fig. 8.1.

Now, referring to Fig. 14.10, we see that a circuit curve which cuts the axis at a shallow angle (a high-impedance circuit curve) will approach or be tangent to the electronic curve at a point where the second derivative is small, while a steep (low impedance) circuit curve will approach the electronic curve at a point where the second derivative is high. This fits in with the idea that a high impedance should give a high gain and a low impedance should give a low gain.

CHAPTER XV

MAGNETRON AMPLIFIER

Synopsis of Chapter

THE HIGH EFFICIENCY of the magnetron oscillator is attributed to motion of the electrons toward the anode (toward a region of higher d-c potential) at high r-f levels. Thus, an electron's loss of energy to the r-f field is made up, not by a slowing-down of its motion in the direction of wave propagation, but by abstraction of energy from the d-c field.[1]

Warnecke and Guenard[2] have published pictures of magnetron amplifiers and Brossart and Doehler have discussed the theory of such devices.[3]

No attempt will be made here to analyze the large-signal behavior of a magnetron amplifier or even to treat the small-signal theory extensively. However, as the device is very closely related to conventional traveling-wave tubes, it seems of some interest to illustrate its operation by a simple small-signal analysis.

The case analyzed is indicated in Fig. 15.1. A narrow beam of electrons flows in the $+z$ direction, constituting a current I_0. There is a magnetic field of strength B normal to the plane of the paper (in the x direction), and a d-c electric field in the y direction. The beam flows near to a circuit which propagates a slow wave. Fig. 15.3, which shows a finned structure opposed to a conducting plane and held positive with respect to it, gives an idea of a physical realization of such a device. The electron stream could come from a cathode held at some potential intermediate between that of the finned structure and that of the plane. In any event, in the analysis the electrons are assumed to have such an initial d-c velocity and direction as to make them travel in a straight line, the magnetic and electric forces just cancelling.

The circuit equation developed in Chapter XIII in connection with transverse motions of electrons is used. Together with an appropriate ballistical equation, this leads to a fifth degree equation for Γ.

[1] For an understanding of the high-level behavior of magnetrons the reader is referred to: J. B. Fisk, H. D. Hagstrum and P. L. Hartman, "The Magnetron as a Generator of Centimeter Waves," *Bell System Technical Journal*, Vol. XXV, April 1946.
"Microwave Magnetrons" edited by George B. Collins, McGraw-Hill, 1948.

[2] R. Warnecke and P. Guenard, "Sur L'Aide Que Peuvent Apporter en Television Quelques Recentes Conceptions Concernant Les Tubes Electroniques Pour Ultra-Hautes Frequences," *Annales de Radioelectricite*, Vol. III, pp. 259–280, October 1948.

[3] J. Brossart and O. Doehler, "Sur les Proprietes des Tubes a Champ Magnetique Constant: les Tubes a Propagation D'Onde a Champ Magnetique," *Annales de Radioelectricite*, Vol. III, pp. 328–338, October 1948.

The nature of this equation indicates that gain may be possible in two ranges of parameters. One is that in which the electron velocity is near to or equal to (as, (15.25)) the circuit phase velocity. In this case there is gain provided that the transverse component of a-c electric field is not zero, and provided that it is related to the longitudinal component as it is for the circuit of Fig. 15.3. It seems likely that this corresponds most nearly to usual magnetron operation.

The other interesting range of parameters is that near

$$\beta_e/\beta_1 = 1 - \beta_m/\beta_1 \tag{15.31}$$

Here β_e refers to the electrons, β_1 to the circuit and β_m is the cyclotron frequency divided by the electron velocity. When (15.31) holds, there is gain whenever the parameter α, which specifies the ratio of the transverse to the longitudinal fields, is not $+1$. For the circuit of Fig. 15.3, α approaccs $+1$ near the fins if the separation between the fins and the plane is great enough in terms of the wavelength. However, α can be made negative near the fins

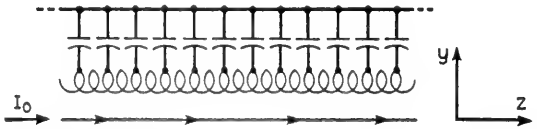

Fig. 15.1—In a magnetron amplifier a narrow electron stream travels in crossed electric and magnetic fields close to a wave transmission circuit.

if the potential of the fins is made negative compared with that of the plane, and the electrons are made to move in the opposite direction.

In either range of parameters, the gain of the increasing wave in db per wavelength is proportional to the square root of the current rather than to the cube root of the current. This means a lower gain than for an ordinary traveling-wave tube with the same circuit and current.

Increasing and decreasing waves with a negative phase velocity are possible when the magnetic field is great enough.

15.1 Circuit Equation

The circuit equation will be the same as that used in Chapter XIII, that is,

$$V = \frac{-j\omega\Gamma_1 K(\Phi\rho - (I_0/u_0)\Phi' y)}{(\Gamma^2 - \Gamma_1^2)} \tag{13.10}$$

It will be assumed that the voltage is given by

$$\Phi = (Ae^{-j\Gamma y} + Be^{j\Gamma y}) \tag{15.1}$$

so that

$$\Phi' V = -jV(Ae^{-j\Gamma y} - Be^{j\Gamma y}) \tag{15.2}$$

Atany$y$ — position we can write

$$\Phi' V = -j\Gamma\alpha\Phi V \tag{15.3}$$

$$\alpha = \frac{Ae^{-j\Gamma y} - Be^{j\Gamma y}}{Ae^{-j\Gamma y} + Be^{j\Gamma y}} \tag{15.4}$$

If Γ is purely imaginary, α is purely real, and as Γ will have only a small real component, α will be considered as a real number. We see that α can range from $+\infty$ to $-\infty$. For instance, consider a circuit consisting of opposed two-dimensional slotted members as shown in Fig. 15.2. For a field with a cosh distribution in the y direction, α is positive above the axis, zero on the axis and negative below the axis. For a field having a sinh distribution in the y

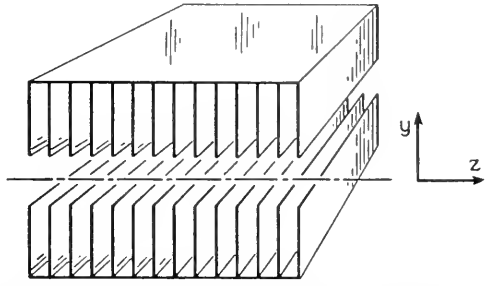

Fig. 15.2—If the circuit is as shown, the ratio between longitudinal and transverse field will be different in sign above and below the axis. This can have an important effect on th operation of the amplifier.

direction, α is infinite on the axis, positive above the axis and negative below the axis.

We find then, that, (13.10) becomes

$$V = \frac{-j\omega\Gamma_1\Phi K(\rho + j\alpha(I_0/u_0)\Gamma y)}{\Gamma^2 - \Gamma_1^2} \tag{15.5}$$

15.2 Ballistic Equations

The d-c electric field in the y direction will be taken as $-E_0$. Thus

$$\frac{d\dot{y}}{dt} = \eta\left[E_0 + \frac{\partial(\Phi V)}{\partial y} - B(z + u_0)\right] \tag{15.6}$$

In order to maintain a rectilinear unperturbed path

$$E_0 = Bu_0 \tag{15.7}$$

so that (15.6) becomes

$$\frac{d\dot{y}}{dt} = \eta\frac{\partial(\Phi V)}{\partial y} - \eta B^z \tag{15.8}$$

Following the usual procedure, we obtain

$$\dot{y} = \frac{-j\eta\Gamma\alpha\Phi V - \eta B^z}{u_0(j\beta_e - \Gamma)} \tag{15.9}$$

We have also

$$\begin{aligned} \frac{d\dot{z}}{dt} &= \eta \frac{\partial \Phi V}{\partial x} + \eta B\dot{y} \\ \dot{z} &= \frac{-\eta\Gamma\Phi V + \eta B\dot{y}}{u_0(j\beta_e - \Gamma)} \end{aligned} \tag{15.10}$$

From (15.9) and (15.10) we obtain

$$z = \frac{-\eta\Gamma\Phi V[(j\beta_e - \Gamma) + j\alpha\beta_m]}{u_0[(j\beta_e - \Gamma)^2 + \beta_m^2]} \tag{15.11}$$

where

$$\beta_m = \omega_m/u_0 \tag{15.12}$$

$$\omega_m = \eta B \tag{15.13}$$

Here ω_m is the cyclotron radian frequency.

As before, we have

$$\rho = \frac{\Gamma\rho_0\dot{z}}{u_0(j\beta_e - \Gamma)} \tag{15.14}$$

whence

$$\rho = \frac{\Gamma^2\eta I_0\Phi V[(j\beta_e - \Gamma) + j\alpha\beta_m]}{u_0^3(j\beta_e - \Gamma)[(j\beta_e - \Gamma)^2 + \beta_m^2]} \tag{15.15}$$

We can also solve (15.9) and (15.10) for $\dot{y}$

$$\dot{y} = \frac{-j\eta\Gamma\Phi V[\alpha(j\beta_e - \Gamma) + j\beta_m]}{u_0[(j\beta_e - \Gamma)^2 + \beta_m^2]}. \tag{15.16}$$

Now, to the first order

$$\begin{aligned} \dot{y} &= \frac{\partial y}{\partial t} + u_0 \frac{\partial y}{\partial z} \\ y &= \frac{\dot{y}}{u_0(j\beta_e - \Gamma)} \end{aligned} \tag{15.17}$$

and from (15.16) and (15.17)

$$y = \frac{-j\eta\Gamma\Phi V[\alpha(j\beta_e - \Gamma) + j\beta_m]}{u_0^2(j\beta_e - \Gamma)[(j\beta_e - \Gamma)^2 + \beta_m^2]}. \tag{15.18}$$

If we use (15.15) and (15.18) in connection with (15.5) we obtain

$$\Gamma^2 - \Gamma_1^2 = \frac{-j\beta_e \Gamma_1 \Gamma^2[(j\beta_e - \Gamma) + 2j[\alpha/(1+\alpha^2)]\beta_m]H^2}{(j\beta_e - \Gamma)[(j\beta_e - \Gamma)^2 + \beta_m^2]} \tag{15.19}$$

$$H^2 = \frac{(1+\alpha^2)\Phi^2 K I_0}{2V_0}. \tag{15.20}$$

Now let

$$-\Gamma_1 = -j\beta_1 \tag{15.21}$$

$$-\Gamma = -j\beta_1(1+p) \tag{15.22}$$

If we assume

$$p \ll 1 \tag{15.23}$$

and neglect p in sums in comparison with unity, we obtain

$$p(\beta_e/\beta_1 - 1 - p)[(\beta_e/\beta_1 - 1 - p)^2 - (\beta_m/\beta_1)^2] = -\frac{\beta_e}{2\beta_1}\left[(\beta_e/\beta_1 - 1 - p) + \frac{2\alpha\beta_m}{(1+\alpha^2)\beta_1}\right]H^2. \tag{15.24}$$

We are particularly interested in conditions which lead to an imaginary value of p which is as large as possible. We will obtain such large values of p when one of the factors multiplying p on the left-hand side of (15.24) is small. There are two possibilities. One is that the first factor is small. We explore this by assuming

$$\beta_e/\beta_1 - 1 = 0 \tag{15.25}$$

$$p^2\left(p^2 - \frac{\beta_m^2}{\beta_1^2}\right) = (1/2)\left(-p + \frac{2\alpha\beta_m}{(1+\alpha^2)\beta_1}\right)H^2, \tag{15.26}$$

If p is very small, we can write approximately

$$-p^2\frac{\beta_m^2}{\beta_1^2} = \frac{\alpha}{(1+\alpha^2)}\frac{\beta_m}{\beta_1}H^2$$
$$p = \pm j[\alpha/(1+\alpha^2)]^{1/2}(\beta_1/\beta_m)^{1/2}H \tag{15.27}$$

We see that p goes to zero if $\alpha = 0$ and is real if α is negative. If we consider what this means circuit-wise, we see that there will be gain with the d-c voltage applied between a circuit and a conducting plane as shown in Fig. 15.3.

Another possible condition in the neighborhood of which p is relatively large is

$$\beta_e/\beta_1 - 1 = \pm\,\beta_m/\beta_1 \tag{15.28}$$

In this case

$$p(\pm\beta_m/\beta_1 - p)(\mp 2(\beta_m/\beta_1)p + p^2) = -\left(1 \pm \frac{\beta_m}{\beta_1}\right)\left[(\pm\beta_m/\beta_1 - p) + \frac{2\alpha\beta_m}{(1+\alpha^2)\beta_1}\right] H^2. \tag{15.29}$$

As p is small, we write approximately

$$p^2 = \pm\frac{1}{4}\frac{(1 \pm \alpha)^2}{1 + \alpha^2}\left(\frac{\beta_1}{\beta_m} \pm 1\right) H^2. \tag{15.30}$$

We see that we obtain an imaginary value of p only for the $-$ sign in (15.28) that is, if

$$\beta_e/\beta_1 = 1 - \beta_m/\beta_1 \tag{15.31}$$

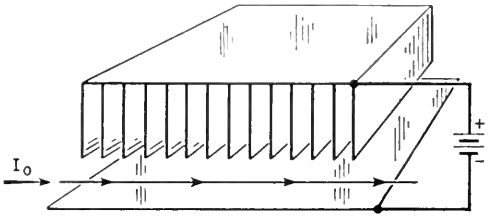

Fig. 15.3—The usual arrangement is to have the finned structure positive and opposed to a conducting plane.

In this case

$$p = \pm j\tfrac{1}{2}[(1 - \alpha)/(1 + \alpha)^{1/2}](\beta_e/\beta_m)^{1/2} H. \tag{15.32}$$

In this case we obtain gain for any value of α smaller than unity. We note that $\alpha = 1$ is the value α assumes far from the axis in a two-dimensional system of the sort illustrated in Fig. 15.2, for either a cosh or a sinh distribution in the $+y$ direction.

The assumption of $-\Gamma = -j\beta_1(1 + p)$ in (15.22) will give forward $(+z)$ traveling-waves only. In order to investigate backward traveling-waves, we must assume

$$-\Gamma = +j\beta_1(1 + p) \tag{15.33}$$

where again p is considered a small number. If we use this in (15.19), we obtain

$$p\left(\frac{\beta_e}{\beta_1} + 1 + p\right)\left[\left(\frac{\beta_e}{\beta_1} + 1 + p\right)^2 - \frac{\beta_m^2}{\beta_1^2}\right] = -\frac{1}{2}\frac{\beta_e}{\beta_1}\left[\left(\frac{\beta_e}{\beta_1} + 1 + p\right) + \frac{2\alpha\beta_m}{(1+\alpha^2)\beta_1}\right] H^2. \tag{15.34}$$

As before we look for solutions for p where the terms multiplying p on the left are small. The only vanishing consistent with positive values of β_e and β_1 is obtained for

$$\frac{\beta_e}{\beta_1} + 1 = +\frac{\beta_m}{\beta_1}. \tag{15.35}$$

Under this condition (15.34) yields for p

$$p = \pm j \frac{1}{2} \frac{(1 + \alpha)}{(1 + \alpha^2)^{1/2}} \left(\frac{\beta_e}{\beta_m}\right)^{1/2} H. \tag{15.36}$$

Thus we can obtain backward-increasing backward-traveling waves for all values of α except $\alpha = -1$. For the situation shown in Fig. 15.3, with a backward wave, α is always negative, approaching -1 at large distances from the plane electrode, so that the gain is identical with that given by (15.32).

We note that (15.27), (15.32) and (15.36) show that p is proportional to the product of current times impedance divided by voltage to the $\frac{1}{2}$ power, while, in the case of the usual traveling-wave tube, this small quantity occurs to the $\frac{1}{3}$ power. The $\frac{1}{3}$ power of a small quantity is larger than the $\frac{1}{2}$ power; and, hence for a given circuit impedance, current and voltage, the gain of the magnetron amplifier will be somewhat less than the gain of a conventional traveling-wave tube.

CHAPTER XVI

DOUBLE-STREAM AMPLIFIERS

Synopsis of Chapter

IN TRAVELING-WAVE TUBES, it is desirable to have the electrons flow very close to the metal circuit elements, where the radio-frequency field of the circuit is strong, in order to obtain satisfactory amplification. It is, however, difficult to confine the electron flow close to metal circuit elements without an interception of electrons, which entails both loss of efficiency and heating of the circuit elements. This latter may be extremely objectionable at very short wavelengths for which circuit elements are small and fragile.

In the double-stream amplifier the gain is not obtained through the interaction of electrons with the field of electromagnetic resonators, helices or other circuits. Instead, an electron flow consisting of two streams of electrons having different average velocities is used. When the currents or charge densities of the two streams are sufficient, the streams interact so as to give an increasing wave.[1,2,3,4] Electromagnetic circuits may be used to impress a signal on the electron flow, or to produce an electromagnetic output by means of the amplified signal present in the electron flow. The amplification, however, takes place in the electron flow itself, and is the result of what may be termed an electromechanical interaction.[5]

While small magnetic fields are necessarily present because of the motions of the electrons, these do not play an important part in the amplification. The important factors in the interaction are the electric field, which stores energy and acts on the electrons, and the electrons themselves. The charge of the electrons produces the electric field; the mass of the electrons, and their kinetic energy, serve much as do inductance and magnetic stored energy in electromagnetic propagation.

[1] J. R. Pierce and W. B. Hebenstreit, "A New Type of High-Frequency Amplifier," *B.S.T.J.*, Vol. 28, pp. 33–51, January 1949.

[2] A. V. Hollenberg, "Experimental Observation of Amplification by Interaction between Two Electron Streams," *B.S.T.J.*, Vol. 28, pp. 52–58, January 1949.

[3] A. V. Haeff, "The Electron-Wave Tube—A Novel Method of Generation and Amplification of Microwave Energy," *Proc. IRE*, Vol. 37, pp. 4–10, January 1949.

[4] L. S. Nergaard, "Analysis of a Simple Model of a Two-Beam Growing-Wave Tube," *R.C.A. Review*, Vol. 9, pp. 585–601, December 1948

[5] Some similar electromechanical waves are described in papers by J. R. Pierce, "Possible Fluctuations in Electron Streams Due to Ions," *Jour. App. Phys.*, Vol. 19, pp. 231–236, March 1948, and "Increasing Space-Charge Waves," *Jour. App. Phys.*, Vol. 20 pp. 1060–1066, Nov. 1949.

By this sort of interaction, a traveling wave which increases as it travels, i.e., a traveling wave of negative attenuation, may be produced. To start such a wave, the electron flow may be made to pass through a resonator or a short length of helix excited by the input signal. Once initiated, the wave grows exponentially in amplitude until the electron flow is terminated or until non-linearities limit the amplitude. An amplified output can be obtained by allowing the electron flow to act on a resonantor, helix or other output circuit at a point far enough removed from the input circuit to give the desired gain.

In general, for a given geometry there is a limiting value of current below which there is no increasing wave. For completely intermingled electron streams, the gain rises toward an asymptotic limit as the current is increased beyond this value. The ordinate of Fig. 16.3 is proportional to gain and the abscissa to current.

When the electron streams are separated, the gain first rises and then falls as the current is increased. This effect, and also the magnitude of the increasing wave set up by velocity modulating the electron streams, have been discussed in the literature.[6]

Double-stream amplifiers have several advantages. Because the electrons interact with one another, the electron flow need not pass extremely close to complicated circuit elements. This is particularly advantageous at very short wavelengths. Further, if we make the distance of electron flow between the input and output circuits long enough, amplification can be obtained even though the input and output circuits have very low impedance or poor coupling to the electron flow. Even though the region of amplification is long, there is no need to maintain a close synchronism between an electron velocity and a circuit wave velocity, as there is in the usual traveling-wave tube.

16.1 Simple Theory of Double-Stream Amplifiers

For simplicity we will assume that the flow consists of coincident streams of electrons of d-c velocities u_1 and u_2 in the z direction. It will be assumed that there is no electron motion normal to the z direction. M.K.S. units will be used.

It turns out to be convenient to express variation in the z direction as

$$\exp -j\beta z$$

rather than as

$$\exp -\Gamma z$$

[6] J. R. Pierce, "Double-Stream Amplifiers," *Proc. I.R.E.*, Vol. 37, pp. 980–985, Sept. 1949.

as we have done previously. This merely means letting

$$\Gamma = j\beta \tag{16.1}$$

The following nomenclature will be used

J_1 , J_2 d-c current densities
u_1 , u_2 d-c velocities
ρ_{01} , ρ_{02} d-c charge densities

$$\rho_{01} = -J_1/u_1 \text{ , } \rho_{02} = -J_2/u_2$$

ρ_1 , ρ_2 a-c charge densities
v_1 , v_2 a-c velocities
V_1 , V_2 d-c voltages with respect to the cathodes
V a-c potential
$\beta_1 = \omega/u_1$, $\beta_2 = \omega/u_2$

From (2.22) and (2.18) we obtain

$$\rho_1 = \frac{\eta J_1 \beta^2 V}{u_1^3(\beta_1 - \beta)^2} \tag{16.2}$$

and

$$\rho_2 = \frac{\eta J_2 \beta^2 V}{u_2^3(\beta_2 - \beta)^2}. \tag{16.3}$$

It will be convenient to call the fractional velocity separation b, so that

$$b = \frac{2(u_1 - u_2)}{u_1 + u_2}. \tag{16.4}$$

It will also be convenient to define a sort of mean velocity u_0

$$u_0 = \frac{2u_1 u_2}{u_1 + u_2}. \tag{16.5}$$

We may also let V_0 be the potential drop specifying a velocity u_0 , so that

$$u_0 = \sqrt{2\eta V_0} \tag{16.6}$$

It is further convenient to define a phase constant based on u_0

$$\beta_0 = \frac{\omega}{u_0}. \tag{16.7}$$

We see from (16.4), (16.5) and (16.6) that

$$\beta_1 = \beta_0(1 - b/2) \tag{16.8}$$

$$\beta_2 = \beta_0(1 + b/2) \tag{16.9}$$

We shall treat only a special case, that in which

$$\frac{J_1}{u_1^3} = \frac{J_2}{u_2^3} = \frac{J_0}{u_0^3}. \tag{16.10}$$

Here J_0 is a sort of mean current which, together with u_0, specifies the ratios J_1/u_1^3 and J_2/u_2^3, which appear in (4) and (5).

In terms of these new quantities, the expression for the total a-c charge density ρ is, from (16.2) and (16.3) and (16.6)

$$\rho = \rho_1 + \rho_2 = \frac{J_0\beta^2}{2u_0V_0} \cdot \left[\frac{1}{\left[\beta_e\left(1 - \frac{b}{2}\right) - \beta\right]^2} + \frac{1}{\left[\beta_e\left(1 + \frac{b}{2}\right) - \beta\right]^2}\right] V. \tag{16.11}$$

Equation (16.11) is a *ballistic* equation telling what charge density ρ is produced when the flow is bunched by a voltage V. To solve our problem, that is, to solve for the phase constant β, we must associate (16.11) with a *circuit* equation which tells us what voltage V the charge density produces. We assume that the electron flow takes place in a tube too narrow to propagate a wave of the frequency considered. Further, we assume that the wave velocity is much smaller than the velocity of light. Under these circumstances the circuit problem is essentially an electrostatic problem. The a-c voltage will be of the same sign as, and in phase with the a-c charge density ρ. In other words the "circuit effect" is purely capacitive.

Let us assume at first that the electron stream is very narrow compared with the tube through which it flows, so that V may be assumed to be constant over its cross section. We can easily obtain the relation between V and ρ in two extreme cases. If the wavelength in the stream is very short (β large), so that transverse a-c fields are negligible, then, from Poisson's equation, we have

$$\rho = -\epsilon \frac{\partial^2 V}{\partial z^2}$$
$$\rho = \epsilon\beta^2 V \tag{16.12}$$

If, on the other hand, the wavelength is long compared with the tube radius (β small) so that the fields are chiefly transverse, the lines of force running from the beam outward to the surrounding tube, we may write

$$\rho = CV \tag{16.13}$$

Here C is a constant expressing the capacitance per unit length between the region occupied by the electron flow and the tube wall.

We see from (16.12) and (16.13) that, if we plot ρ/V vs. β/β_e for real values of β, ρ/V will be constant for small values of β and will rise as β^2 for large values of β, approximately as shown in Fig. 16.1.

Now, we have assumed that the charge is produced by the action of the voltage, according to the ballistical equation (16.11). This relation is plotted in Fig. 2, for a relatively large value of J_0/u_0V_0 (curve 1) and for a smaller value of J_0/u_0V_0 (curve 2). There are poles at $\beta/\beta_e = 1 \pm \frac{b}{2}$, and a minimum between the poles. The height of the minimum increases as J_0/u_0V_0 is increased.

A circuit curve similar to that of Fig. 16.1 is also plotted on Fig. 16.2. We see that for the small-current case (curve 2) there are four intersections, giving *four real* values of β and hence *four unattenuated* waves. However, for

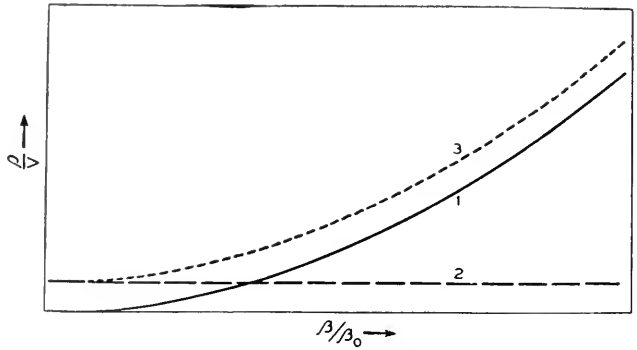

Fig. 16.1—Circuit curves, in which the ordinate is proportional to the ratio of the charge per unit length to the voltage which it produces. Curve 1 is for an infinitely broad beam; curve 2 is for a narrow beam in a narrow tube. Curve 3 is the sum of 1 and 2, and approximates an actual curve.

the larger current (curve 1) there are only two intersections and hence two unattenuated waves. The two additional values of β satisfying both the circuit equation and the ballistical equation are complex conjugates, and represent waves traveling at the same speed, but with equal positive negative attenuations.

Thus we deduce that, as the current densities in the electron streams are raised, a wave with negative attenuation appears for current densities above a certain critical value.

We can learn a little more about these waves by assuming an approximate expression for the circuit curve of Fig. 1. Let us merely assume that over the range of interest (near $\beta/\beta_e = 1$) we can use

$$\rho = \alpha^2\epsilon\beta^2 V \tag{16.14}$$

Here α^2 is a factor greater than unity, which merely expresses the fact that the charge density corresponding to a given voltage is somewhat greater than if there were field in the z direction only for which equation (16.12) is valid. Combining (16.14) with (16.11) we obtain

$$\frac{1}{\left(\beta_e\left(1-\frac{b}{2}\right)-\beta\right)^2}+\frac{1}{\left(\beta_e\left(1+\frac{b}{2}\right)-\beta\right)^2}=\frac{1}{\beta_e^2 U^2} \tag{16.15}$$

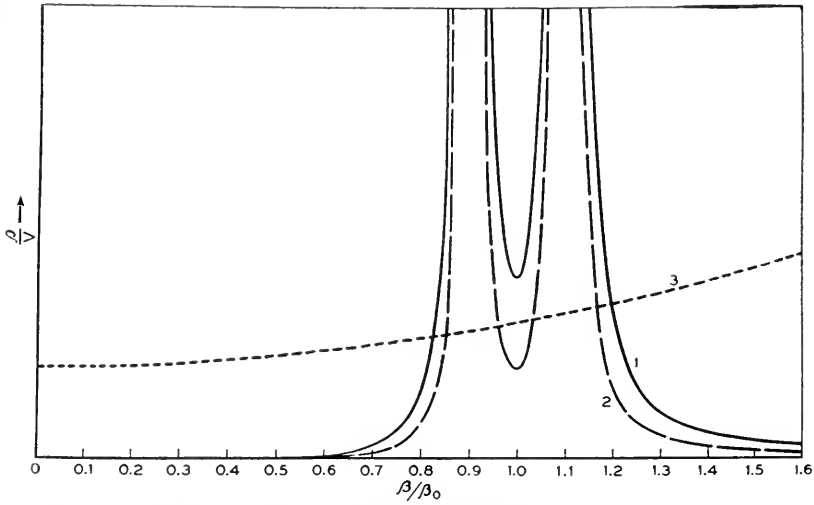

Fig. 16.2—This shows a circuit curve, 3, and two electronic curves which give the sum of the charge densities of the two streams divided by the voltage which bunches them. With curve 2, there will be four unattenuated waves. With curve 1, which is for a higher current density than curve 2, there are two unattenuated waves, an increasing wave and a decreasing wave.

where

$$U^2=\frac{J_0}{2\alpha^2\epsilon\beta_e^2 u_0 V_0}. \tag{16.16}$$

In solving (16.15) it is most convenient to represent β in terms of β_e and a new variable h

$$\beta=\beta_e(1+h) \tag{16.17}$$

Thus, (16.15) becomes

$$\frac{1}{\left(h-\frac{b}{2}\right)^2}+\frac{1}{\left(h+\frac{b}{2}\right)^2}=\frac{1}{U^2}. \tag{16.18}$$

Solving for h, we obtain

$$h = \pm\left(\frac{b}{2}\right)\left[\left(\frac{2U}{b}\right)^2 + 1 \pm \left(\frac{2U}{b}\right)\sqrt{\left(\frac{2U}{b}\right)^2 + 4}\right]^{1/2}. \quad (16.19)$$

The positive sign inside of the brackets always gives a real value of h and hence unattenuated waves. The negative sign inside the brackets gives unattenuated waves for small values of U/b. However, when

$$\left(\frac{U}{b}\right)^2 > \frac{1}{8} \quad (16.20)$$

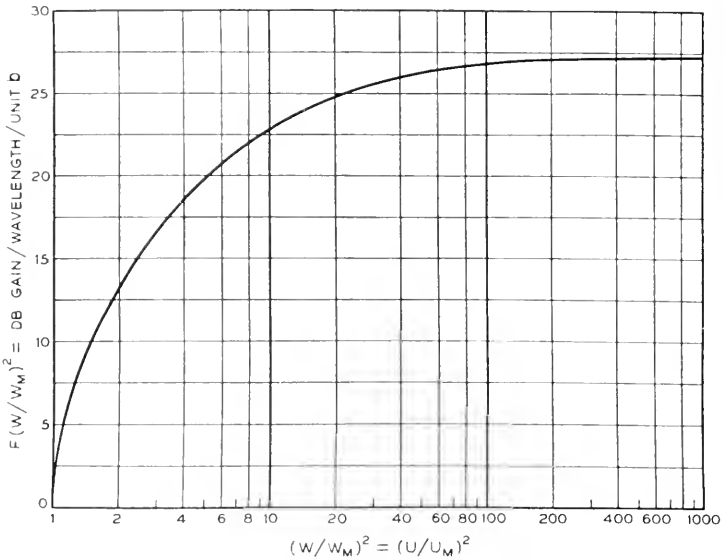

Fig. 16.3—The abscissa is proportional to d-c current. As the current is increased, the gain in db per wavelength approaches 27.3b, where b is the fractional separation in velocity. If the two electron streams are separated physically, the gain is lower and first rises and then falls as the current is increased.

there are two waves with a phase constant β_e and with equal and opposite attenuation constants.

Suppose we let U_M be the minimum value of U for which there is gain. From (16.20)

$$U_M^2 = b^2/8 \quad (16.21)$$

From (16.19) we have, for the increasing wave,

$$h = jb\left[\frac{U}{\sqrt{2}\,U_M}\sqrt{2\left(\frac{U}{U_M}\right)^2 + 1} - \left(\frac{U}{U_M}\right)^2 - 1\right]^{1/2}. \quad (16.22)$$

The gain in db/wavelength is

$$\begin{aligned} \text{db/wavelength} &= 20(2\pi)\log_{10}e^{|h|} \\ &= 54.6\,|\,h\,| \end{aligned} \tag{16.23}$$

We see that, by means of (16.22) and (16.23), we can plot db/wavelength per unit b vs. $(U/U_M)^2$. This is plotted in Fig. 16.3. Because U^2 is proportional to current, the variable $(U/U_M)^2$ is the ratio of the actual current to the current which will just give an increasing wave. If we know this ratio, we can obtain the gain in db/wavelength by multiplying the corresponding ordinate from Fig. 16.3 by b.

We see that, as the current is increased, the gain per wavelength at first rises rapidly and then rises more slowly, approaching a value $27.3b$ db/wavelength for very large values of $(U/U_M)^2$.

We now have some idea of the variation of gain per wavelength with velocity separation b and with current $(U/U_M)^2$. A more complete theory requires the evaluation of the lower limiting current for gain (or of U_M^2) in terms of physical dimensions and an investigation of the boundary conditions to show how strong an increasing wave is set up by a given input signal.[1, 6]

16.2 Further Considerations

There are a number of points to be brought out concerning double-stream amplifiers. Analysis shows[6] that any physical separation of the electron streams has a very serious effect in reducing gain. Thus, it is desirable to intermingle the streams thoroughly if possible.

If the electron streams have a fractional velocity spread due to space charge which is comparable with the deliberately imposed spread b, we may expect a reduction in gain.

Haeff[3] describes a single-stream tube and attributes its gain to the space-charge spread in velocities. In his analysis of this tube he divides the beam into a high and a low velocity portion, and assigns the mean velocity to each. This is not a valid approximation.

Analysis indicates that a multiply-peaked distribution of current with velocity is necessary for the existent increasing waves, and gain in a "single stream" of electrons is still something of a mystery.

CHAPTER XVII

CONCLUSION

ALTHOUGH THIS BOOK contains some descriptive material concerning high-level behavior, it is primarily a treatment of the linearized or low-level behavior of traveling-wave tubes and of some related devices. In the case of traveling-wave tubes with longitudinal motion of electrons only, the treatment is fairly extended. In the discussions of transverse fields, magnetron amplifiers and double-stream amplifiers, it amounts to little more than an introduction.

One problem to which the material presented lends itself is the calculation of gain of longitudinal-field traveling-wave tubes. To this end, a summary of gain calculation is included as Appendix VII.

Further design information can be worked out as, for instance, exact gain curves at low gain with lumped or distributed loss, perhaps taking the space-charge parameter QC into account, or, a more extended analysis concerning noise figure.

The material in the book may be regarded from another point of view as an introduction, through the treatment of what are really very simple cases, to the high-frequency electronics of electron streams. That is, the reader may use the book merely to learn how to tackle new problems. There are many of these.

One serious problem is that of extending the non-linear theory of the traveling-wave tube. For one thing, it would be desirable to include the effects of loss and space charge. Certainly, a matter worthy of careful investigation is the possibility of increasing efficiency by the use of a circuit in which the phase velocity decreases near the output end. Nordsieck's work can be a guide in such endeavors.

Even linear theory excluding the effects of thermal velocities could profitably be extended, especially to disclose the comparative behavior of narrow electron beams and of broad beams, both those confined by a magnetic field, in which transverse d-c velocities are negligible and in which space charge causes a lowering of axial velocity toward the center of the beam, and also those in which transverse a-c velocities are allowed, especially the Brillouin-type flow, in which the d-c axial velocity is constant across the beam, but electrons have an angular velocity proportional to radius.

Further problems include the extension of the theory of magnetron amplifiers and of double-stream amplifiers to a scope comparable with that of the

theory of conventional traveling-wave tubes. The question of velocity distribution across the beam is particularly important in double-stream amplifiers, whose very operation depends on such a distribution, and it is important that the properties of various kinds of distribution be investigated.

Finally, there is no reason to suspect that the simple tubes described do not have undiscovered relatives of considerable value. Perhaps diligent work will uncover them.

BIBLIOGRAPHY

1946

Barton, M. A. Traveling wave tubes, *Radio*, v. 30, pp. 11–13, 30–32, Aug., 1946.

Blanc-Lapierre, A. and Lapostolle, P. Contribution à l'étude des amplificateurs à ondes progressives, *Ann. des Telecomm.*, v. 1, pp. 283–302, Dec., 1946.

Kompfner, R. Traveling wave valve—new amplifier for centimetric wavelengths. *Wireless World*, v. 52, pp. 369–372, Nov., 1946.

Pierce, J. R. Beam traveling-wave tube, *Bell Lab. Record*, v. 24, pp. 439–442, Dec., 1946.

1947

Bernier, J. Essai de théorie du tube électronique à propagation d'onde, *Ann. de Radioélec.*, v. 2, pp. 87–101, Jan., 1947. *Onde Élec.*, v. 27, pp. 231–243, June, 1947.

Blanc-Lapierre, A., Lapostolle, P., Voge, J. P., and Wallauschek, R. Sur la théorie des amplificateurs à ondes progressives, *Onde Élec.*, v. 27, pp. 194–202, May, 1947.

Kompfner, R. Traveling-wave tuge as amplifier at microwaves, *I.R.E., Proc.*, v. 35, pp. 124–127, Feb., 1947.

Kompfner, R. Traveling-wave tube—centimetre-wave amplifier, *Wireless Engr.*, v. 24, pp. 255–266, Sept., 1947.

Pierce, J. R. Theory of the beam-type traveling-wave tube, *I.R.E., Proc.*, v. 35, pp. 111–123, Feb., 1947.

Pierce, J. R. and Field, L. M. Traveling-wave tubes, *I.R.E., Proc.*, v. 35, pp. 108–111, Feb., 1947.

Roubine, E. Sur le circuit à hélice utilisé dans le tube à ondes progressives, *Onde Élec.*, v. 27, pp. 203–208, May, 1947.

Shulman, C. and Heagy, M. S. Small-signal analysis of traveling-wave tube, *R.C.A. Rev.*, v. 8, pp. 585–611, Dec., 1947.

1948

Brillouin, L. Wave and electrons traveling together—a comparison between traveling wave tubes and linear accelerators, *Phys. Rev.*, v. 74, pp. 90–92, July 1, 1948.

Brossart, J. and Doehler, O. Sur les propriétés des tubes à champ magnétique constant. Les tubes à propagation d'onde à champ magnétique, *Ann. de Radioélec.*, v. 3, pp. 328–338, Oct., 1948.

Cutler, C. C. Experimental determination of helical-wave properties, *I.R.E., Proc.*, v. 36, pp. 230–233, Feb., 1948.

Chu, L. J. and Jackson, J. D. Field theory of traveling-wave tubes, *I.R.E., Proc.*, v. 36, pp. 853–863, July, 1948.

Döehler, O. and Kleen, W. Phénomènes non lineaires dans les tubes à propagation d'onde. *Ann. de Radioélec.*, v. 3, pp. 124–143, Apr., 1948.

Döehler, O. and Kleen, W. Sur l'influence de la charge d'espace dans le tube à propagation d'onde, *Ann. de Radioélec.*, v. 3, pp. 184–188, July, 1948.

Blanc-Lapierre, A., Kuhner, M., Lapostolle, P., Jessel, M. and Wallauschek, R. Étude et réalisation d'amplificateurs à hélice. *Ann. des Telecomm.*, v. 3, pp. 257–308, Aug.–Sept., 1948.

Blanc-Lapierre, A. and Kuhner, M. Réalisation d'amplificateurs à onde progressive à hélice. Résultats généraux, pp. 259–264.

Lapostolle, P. Les phénomènes d'interaction dans le tube à onde progressive, Théorie et vérifications expérimentales, pp. 265–291.

Jessel, M. and Wallauschek, R. Étude expérimentale de la propagation de long d'une ligne à retard en forme d'hélice, pp. 291–299.

Wallauschek, R. Détérmination expérimentale des caractéristiques d'amplificateurs à onde progressive, Résultats obtenus, pp. 300–308.

Lapostolle, P. Étude des diverses ondes susceptibles de se propager dans une ligne en interaction avec un faisceau électronique. Application à la théorie de l'amplificateur à onde progressive, *Ann. des Telecomm.*, v. 3, pp. 57–71, Feb., pp. 85–104, Mar., 1948.

Pierce, J. R. Effect of passive modes in traveling wave tubes, *I.R.E., Proc.*, v. 36, pp. 993–997, Aug., 1948.
Pierce, J. R. Transverse fields in traveling-wave tubes, *Bell Sys. Tech. Jl.*, v. 27, pp. 732–746, Oct., 1948.
Rydbeck, O. E. H. Theory of the traveling-wave tube, *Ericsson Technics*, no. 46, pp. 3–18, 1948.
Tomner, J. S. A. Experimental development of traveling-wave tubes, Acta Polytech., *Elec. Engg.*, v. 1, no. 6, pp. 1–21, 1948.
Nergaard, L. S. Analysis of a simple model of a two-beam growing-wave tube, RCA Rev. vol. 9, pp. 585–601, Dec. 1948.

1949

Doehler, O. and Kleen, W. Influence du vecteur électrique transversal dans la ligne à retard du tube à propagation d'onde, *Ann. de Radioélec.*, v. 4, pp. 76–84, Jan., 1949.
Bruck, L. Comparison des valeurs mesurees pour le gain lineaine du tube à propagation d'onde avec les valeurs indiquées par diverses theories. *Annales de Radioélectricete*, v. IV, pp. 222–232, July, 1949.
Döehler, O. and Kleen, W. Sur le rendement du tube à propagation d'onde. *Annales de Radioélectricete*, v. IV, pp. 216–221, July, 1949.
Döehler, O., Kleen, W. and Palluel, P. Les tubes à propagation d'onde comme oscillateurs à large bande d'accord électronique. *Ann. de Radioélec.*, v. 4, pp. 68–75, Jan., 1949.
Döhler, O. and Kleen, W. Über die Wirkungsweise der "Traveling-Wave" Röhre. *Arch. Elektr. Übertragung*, v. 3, pp. 54–63, Feb., 1949.
Field, L. M. Some slow-wave structures for traveling-wave tubes. *I.R.E., Proc.*, v. 37, pp. 34–40, Jan., 1949.
Guenard, P., Berterattiere, R. and Doehler, O. Amplification par interaction electronque dans des tubes sans circuits. *Annales de Radioélectricete*, v. IV, pp. 171–177, July, 1949.
Laplume, J. Théorie du tube à onde progressive. *Onde Élec.*, v. 29, pp. 66–72, Feb., 1949.
Loshakov, L. N. On the propagation of Waves along a coaxial spiral line in the presence of an electron beam. *Zh. Tech. Fiz.*, vol. 19, pp. 578–595, May, 1949.
Dewey, G. C. A periodic-waveguide traveling-wave amplifier for medium powers. *Proc. N.E.C.* (Chicago), v. 4, p. 253, 1948.
Pierce, J. R. and Hebenstreit, W. B. A new type of high-frequency amplifier, Bell System Technical Journal, v. 28, pp. 33–51, January, 1949.
Haeff, A. V. The electron-wave tube—a novel method of generation and amplification of microwave energy. *Proc. I.R.E.*, v. 37, pp. 4–10, January, 1949.
Hollenberg, A. V. The double-stream amplifier, *Bell Laboratories Record*, v. 27, pp. 290–292, August, 1949.
Guenard, P., Berterottiere, R. and Döehler, O. Amplification by direct electronic interaction in valves without circuits, *Ann. Radioelec.*, v. 4, pp. 171–177, July, 1949.
Rogers, D. C. Traveling-wave amplifier for 6 to 8 centimeters, *Elec. Commun.* (London), v. 26, pp. 144–152, June, 1949.
Field, L. M. Some slow wave structures for traveling-wave tubes, *Proc. I.R.E.*, v. 37, pp. 34–40, January, 1949.
Schnitzer, R. and Weber, D. *Frequenz*, v. 3, pp. 189–196, July, 1949.
Pierce, J. R. Circuits for traveling-wave tubes, *Proc. I.R.E.*, v. 37, pp. 510–515, May, 1949.
Pierce, J. R. and Wax, N. A note on filter-type traveling-wave amplifiers, *Proc. I.R.E.*, v. 37, pp. 622–625, June, 1949.

Technical Publications by Bell System Authors Other Than in the Bell System Technical Journal

*Circuits for Cold Cathode Glow Tubes.** W. A. Depp[1] and W. H. T. Holden.[1] *Elec. Mfg.*, v. 44, pp. 92–97, July, 1949.

Equipment for the Determination of Insulation Resistance at High Humidities. A. T. Chapman.[2] *A.S.T.M. Bull.*, no. 165, pp. 43–45, Apr., 1950.

Twin Relationships in Ingots of Germanium. W. C. Ellis.[1] *Jl. Metals*, v. 188, p. 886, June, 1950.

*Magnetic Cores of Thin Tape Insulated by Cataphoresis.** H. L. B. Gould.[1] *Elec. Engg.*, v. 69, pp. 544–548, June, 1950.

Slip Markings in Chromium. E. S. Greiner.[1] *Jl. Metals*, v. 188, pp. 891–892, June, 1950.

New Porcelain Rod Leak. H. D. Hagstrum[1] and H. W. Weinhart.[1] *Rev. Sci. Instruments*, v. 21, p. 394, Apr., 1950.

Significance of Nonclassical Statistics. R. V. L. Hartley.[1] *Science*, v. 111, pp. 574–576, May 26, 1950.

Comment on Mobility Anomalies in Germanium. G. L. Pearson,[1] J. R. Haynes[1] and W. Shockley.[1] Letter to the editor. *Phys. Rev.*, v. 78, pp. 295–296, May 1, 1950.

Dislocation Models of Crystal Grain Boundaries. W. T. Read[1] and W. Shockley.[1] *Phys. Rev.*, v. 78, pp. 275–289, May 1, 1950.

Zero-Point Vibrations and Superconductivity. J. Bardeen.[1] Letter to the editor. *Phys. Rev.*, v. 79, pp. 167–168, July 1, 1950.

Some Observations on Industrial Research. O. E. Buckley.[1] *Bell Tel. Mag.*, v. 29, pp. 13–24, Spring, 1950.

Properties of Single Crystals of Nickel Ferrite. J. K Galt,[1] B. T. Mathias[1] and J. P. Remeika.[1] Letter to the editor. *Phys. Rev.*, v. 79, pp. 391–392, July 15, 1950.

Photon Yield of Electron-Hole Pairs in Germanium. F. S. Goucher.[1] Letter to the editor. *Phys. Rev.*, v. 78, p. 816, June 15, 1950.

Data on Porcelain Rod Leak J. P. Molnar[1] and C. D. Hartman.[1] *Rev. Sci. Instruments*, v. 21, pp. 394–395, Apr., 1950.

Magnetic Susceptibility of αFe_2O_3 *and* αFe_2O_3 *with Added Titanium.* F. J. Morin.[1] Letter to the editor. *Phys. Rev.*, v. 78, pp. 819–820, June 15, 1950.

* A reprint of this article may be obtained on request to the editor of the B.S.T.J.

[1] B.T.L.

[2] W.E.Co.

Alternating Current Conduction in Ice. E. J. MURPHY.[1] Letter to the editor. *Phys. Rev.*, v. 79, pp. 396–397, July 15, 1950.

Ferromagnetic Resonance in Manganese Ferrite and the Theory of the Ferrites. W. A. YAGER,[1] F. R. MERRITT,[1] C. KITTEL[1] and C. GUILLAUD.[1] Letter to the editor. *Phys. Rev.*, v. 79, p. 181, July 1, 1950.

Conductivity Pulses Induced in Diamond by Alpha Particles. A. J. AHEARN.[1] *AEC, Brookhaven conference report, BNL-C-1 High speed counters and short pulse techniques*, Aug. 14–15, 1947. 1950. p. 7.

Behavior of Resistors at High Frequencies. G. R. ARTHUR[1] and S. E. CHURCH.[1] *T. V. Engg.*, v. 1, pp. 4–7, June, 1950.

Note on "The Application of Vector Analysis to the Wave Equation". R. V. L. HARTLEY.[1] Letter to the editor. *Acoustical Soc. Am. Jl.*, v. 22, p. 511, July, 1950.

*Number 5 Crossbar Dial Telephone Switching System.** F. A. KORN[1] and JAMES G. FERGUSON.[1] *Elec. Engg.*, v. 69, pp. 679–684, Aug., 1950.

Traveling-Wave Tube as a Broad Band Amplifier. J. R. PIERCE.[1] *AEC, Brookhaven conference report, BNL-C-1 High speed counters and short pulse techniques*, Aug. 14–15, 1947. 1950. p. 41.

* A reprint of this article may be obtained on request to the editor of the B.S.T.J.
[1] B.T.L.

Contributors to this Issue

JOHN BARDEEN, University of Wisconsin, B.S. in E.E., 1928; M.S., 1930. Gulf Research and Development Corporation, 1930–33; Princeton University, 1933–35, Ph.D. in Math. Phys., 1936; Junior Fellow, Society of Fellows, Harvard University, 1935–38; Assistant Professor of Physics, University of Minnesota, 1938–41; Prin. Phys., Naval Ordnance Laboratory, 1941–45. Bell Telephone Laboratories, 1945–. Dr. Bardeen is engaged in theoretical problems related to semiconductors.

A. E. BOWEN, Ph.B., Yale University, 1921; Graduate School, Yale University, 1921–24. American Telephone and Telegraph Company, Department of Development and Research, 1924–34. Bell Telephone Laboratories, 1934–42. U. S. Army Air Force, 1942–45. Bell Telephone Laboratories, 1945–48. With the American Telephone and Telegraph Company, Mr. Bowen's work was concerned principally with the inductive coordination of power and communications systems. From 1934 to 1942 he was engaged in work in the ultra-high-frequency field, particularly on hollow waveguides. He became a Major and later a Colonel while serving with the U. S. Army Air Force from 1942 to 1945 on a special mission to Trinidad and subsequently in the Pentagon. After returning to Bell Telephone Laboratories in 1945 he was engaged in the problems of microwave repeater research until his death in 1948.

M. E. HINES, B.S. in Applied Physics, California Institute of Technology, 1940; B.S. in Meteorology, 1941; M.S. in Electrical Engineering, 1946. U. S. Air Force Weather Service, 1941–45. Bell Telephone Laboratories, 1946–. Mr. Hines has been engaged in the development of vacuum tubes.

JACK A. MORTON, B.S. in Electrical Engineering, Wayne University, 1935; M.S.E., University of Michigan, 1936. Bell Telephone Laboratories, 1936–. Mr. Morton joined the Laboratories to work on coaxial cable and microwave amplifier circuit research; during the war he was at first a member of a group engaged in improving the signal-to-noise performance of radar receivers. In 1943 he transferred to the Electronic Development Department to work on microwave tubes for radar and radio relay. Since 1948 he has been Electronic Apparatus Development Engineer responsible for the development of transistors and other semiconductor devices.

WILLIAM W. MUMFORD, B.A., Willamette University, 1930. Bell Telephone Laboratories, 1930–. Mr. Mumford has been engaged in work that is chiefly concerned with ultra-short-wave and microwave radio communication.

J. R. PIERCE, B.S. in Electrical Engineering, California Institute of Technology, 1933; Ph.D., 1936. Bell Telephone Laboratories, 1936–. Dr. Pierce has been engaged in the study of vacuum tubes.

ROBERT M. RYDER, Yale University, B.S. in Physics, 1937; Ph.D., 1940. Bell Telephone Laboratories, 1940–. Dr. Ryder joined the Laboratories to work on microwave amplifier circuits, and during most of the war was a member of a group engaged in studying the signal-to-noise performance of radars. In 1945 he transferred to the Electronic Development Department to work on microwave oscillator and amplifier tubes for radar and radio relay applications. He is now in a group engaged in the development of transistors.

W. VAN ROOSBROECK, A.B., Columbia College, 1934; A.M., Columbia University, 1937. Bell Telephone Laboratories, 1937–. Mr. van Roosbroeck's work at the Laboratories was concerned during the war with carbon-film resistors and infra-red bolometers and, more recently, with the copper oxide rectifier. In 1948 he transferred to the Physical Research Department where he is now engaged in problems of solid-state physics.