REVIEWS OF

MODERN PHYSICS

VOLUME 18, NUMBER 4 OCTOBER, 1946

Microwave Electronics

J. C. SLATER
Massachusetts Institute of Technology, Cambridge, Massachusetts

BY microwave electronics one can refer to the study of electromagnetic fields in regions of the order of a wave-length in dimensions, bounded by reflecting walls, and of the interaction of these fields with electrons, ions, or other forms of space charge. This includes the whole high frequency side of microwave radar; the nature of wave guides and resonant cavities; and the nature of electronic tubes, such as klystrons and magnetrons, in which transit time is of fundamental importance. It also includes such devices as cyclotrons, synchrotrons, linear accelerators, and other devices for the acceleration of charged particles. This article presents some aspects of this field, but is far from a complete treatment. It represents essentially a set of lecture notes for a series of seminars delivered by the writer during the winter of 1945–46. It is hoped that the material eventually will be expanded greatly, into a full-sized book, to be published by D. Van Nostrand Company, in the series of publications from the Bell Telephone Laboratories. Since, however, this publication will not appear for some time, it was felt worth while presenting this abridged and incomplete version, on account of the great present lack of material on the subject.

During the war there was, of course, a very great development in the knowledge of microwaves. The study of wave guides and resonant cavities, originated before the war by Barrow, Chu, Schelkunoff, and others, was carried to a point of great advancement. The klystron, developed before the war by Hansen, the Varians, and others, became much better known and highly perfected. The magnetron oscillator was improved, first in England and then in this country, to the point where it was a generator of microwave power of very great capabilities. All of these developments, particularly as they were carried on at the M. I. T. Radiation Laboratory, will be discussed at length in the great series of books to be issued from that laboratory, and published by the McGraw-Hill Book Company, Inc. Other versions of the same information have been, or will be, contained in publications from the industrial laboratories in various periodicals, such as the *Bell System Technical Journal*. Even these publications, extensive as they will be, represent but a small fraction of the great literature which exists in the form of classified, or formerly classified, reports prepared during the war. With this large program of projected publication, it is worth asking why the present review article, and its future enlargement into a book, are necessary.

The answer would be that the author has tried to introduce into the field a correlation and unity which are perhaps lacking in most of the other work. During the war his work was largely on magnetrons, both at the M. I. T. Radiation Laboratory and at the Bell Telephone Laboratories, of which he was for a time a staff member. It became clear in the early stages of this work that a study merely of the electronics of the magnetron was incomplete and unsatisfactory; it was necessary in addition to take into account the resonant circuit, consisting of resonant cavities and the attached loads, and to consider the reaction of this circuit back on the electronic motions. This in turn led to the development of a circuit theory of resonant cavities, and of the wave guides which form the leads of these cavities, based on the theory of orthogonal functions, and on the expansion of Maxwell's equations in a closed region in terms of such orthogonal functions. This development gives a logically satisfying foundation for the whole of microwave electronics, and at the same time proves to be of great practical value in the design and development of magnetrons. Later application to reflex klystrons has shown that the principles are of wide importance. Work since the war has convinced the writer that in such problems as the linear accelerator the methods are of just as much value. The main purpose of this review article is to present this unified point of view, carrying the application to problems such as the klystron and the magneton only far enough to illustrate the general method. The later amplification in book form will carry these applications much farther, and to a wider variety of problems.

The work presented in the present article, of course, represents contributions from a variety of workers, and no attempt is made to assign credit for it. The second chapter, on wave guides, is to some extent familiar. The orthogonal function development was worked out, not only by the author, but by Bethe in the Radiation Laboratory, and presumably by others. The material of the third chapter, on resonant cavities, was suggested to the author by a treatment given by Condon in 1941, but in a much more incomplete form. At about the same time that the writer was working it out in the Radiation Laboratory, Schwinger was also carrying out very similar expansions in orthogonal functions, and using them for similar purposes, though the two developments were largely independent. Feshbach, of the Massachusetts Institute of Technology, was also working along similar lines. Much of the material of the fourth chapter, on the applications of the theory of resonant cavities, was common knowledge at the Radiation Laboratory, the result of work of Lawson, Rieke, and others, though its formulation in terms of the general theory of resonant cavities, and a number of the specific applications as well, originated with the author. The fifth chapter, application to the electronics of the reflex klystron and the magnetron, likewise represents the ideas of many, both in this country and in England, though the author was associated with all parts of the subject, and wrote extensive reports, much fuller than is suggested here, on both the electronic and circuit phases.

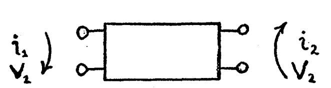

FIG. 1. A four-terminal network.

It seems worth while pointing out that the theory of expansion of solutions of Maxwell's equations in orthogonal functions, which is presented here, can have useful applications in the theory of quantum electrodynamics, as well as in microwaves. The scalar and vector potentials, as well as the fields, can be expanded in these orthogonal functions, and the Lagrangian and Hamiltonian of the field can be set up in terms of these expansion coefficients. By this means we can set up a classical relativistic Hamiltonian theory of the interaction of fields and matter in an arbitrary cavity, which can then be carried over into quantum theory in a manner similar to that of Dirac's radiation theory. This frees that theory from the requirement that the field be expanded in plane waves, and provides a much more general form of expansion. While this does not seem to remove any of the outstanding difficulties in quantum electrodynamics, it yields a new point of view which may be useful. The author hopes to develop this application in a later paper.

Several books have appeared in the last few years, treating the pre-war status of the subject satisfactorily. Among these we list the following:

J. G. Brainerd, G. Koehler, H. J. Reich, and L. F. Woodruff, *Ultra-High Frequency Techniques* (D. Van Nostrand Company, New York, 1942).

S. Ramo and J. R. Whinnery, *Fields and Waves in Modern Radio* (John Wiley and Sons, Inc., New York, 1944).

R. I. Sarbacher and W. A. Edson, *Hyper and Ultrahigh Frequency Engineering* (John Wiley and Sons, Inc., New York, 1943).

S. A. Schelkunoff, *Electromagnetic Waves* (D. Van Nostrand Company, New York, 1943).

J. C. Slater, *Microwave Transmission* (McGraw-Hill Book Company, Inc., New York, 1942).

J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941).

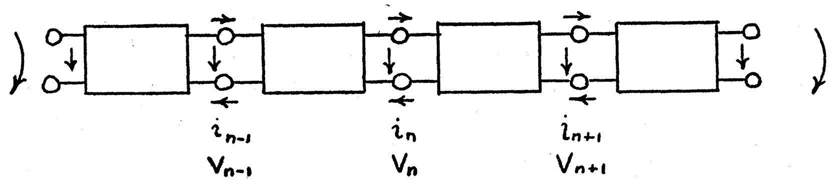

FIG. 2. A transmission line.

The book of Sarbacher and Edson contains a very complete and excellent bibliography of the pre-war literature of the whole field. Since it is so complete, it seems unnecessary to give references to that literature here. Various review and other articles have appeared since the books mentioned above. Among these may be mentioned the following:

E. U. Condon, "Principles of Microwave Radio," Rev. Mod. Phys. **14**, 341 (1942). This article presents a point of view similar to that of the present one in the matter of the oscillations of cavity resonators, but does not carry the application to self-excited oscillators as far as we do in the present article.

J. B. Fisk, H. D. Hagstrum, and P. L. Hartman, "The Magnetron as a Generator of Centimeter Waves," Bell Sys. Tech. J. **25**, 167 (1946). This is an excellent review of the magnetron development during the war, treating both the theoretical and the practical side, and can well be used to supplement the discussion of magnetrons in the present article.

J. R. Pierce, "Reflex Oscillators," Proc. I.R.E. **33**, 112 (1945).

E. L. Ginzton and A. E. Harrison, "Reflex-Klystron Oscillators," Proc. I.R.E. **34**, 97 (1946). These two articles present information on reflex klystrons which supplements the treatment of the present article, but they are not complete reviews in the sense that the article of Fisk, Hagstrum, and Hartman is in its field.

The references which we have just enumerated will be sufficient to allow the reader rather easily to become acquainted with the published literature in the field. The main references, however, should properly be to the unpublished material which is scheduled to come out within a year, principally the Radiation Laboratory Series, and they will contain material far more complete than anything that has appeared in print so far.

I. THE FOUR-TERMINAL NETWORK AND THE TRANSMISSION LINE

1. Definition of the Four-Terminal Network and the Transmission Line

The principles of microwave electronics can be developed entirely without using the theory of lumped constant circuits, and we shall so develop them. Nevertheless, the results are so similar to those of ordinary circuits that a knowledge of the fundamentals of circuit theory is an excellent background for understanding microwave circuits. For that reason we start by discussing a number of theorems in circuit theory. First we consider a fundamental starting point, the theory of the four-terminal network. This is a collection of circuit elements whose nature we do not have to inquire into, except that it is provided with two input and two output terminals. Let V_1, V_2 be the voltages across the two sets of terminals, and i_1, i_2 the currents flowing, the currents being positive when flowing in the direction of the arrows in Fig. 1, and the voltages positive when the arrows point in the direction of decreasing voltage. Then if the network is linear, the voltages will be linear functions of the currents:

$$V_1 = Z_{11}i_1 + Z_{12}i_2,\qquad V_2 = Z_{21}i_1 + Z_{22}i_2, \tag{I.1}$$

where it can be proved that $Z_{12} = Z_{21}$, the so-called reciprocity relation. The Z's are quantities of the nature of impedances. By a transmission line we mean a set of many identical four-terminal networks, connected together. Each one is governed by equations like (I.1). For symmetry, however, it is more convenient to choose the convention of signs differently. As in Fig. 2, we choose all voltages to be positive when the upper terminal is at higher voltage, all currents to be positive when they flow to the right in the upper terminals. That is, the sign of the voltage at the right-hand terminals of a network is reversed with respect to the convention of (I.1), but all other quantities are unchanged. In this case, with the Z's meaning the same as above, we have the following relations between the currents and voltages at the nth and $(n+1)$st terminals:

$$V_n = Z_{11}i_n + Z_{12}i_{n+1}\qquad V_{n+1} = -Z_{12}i_n - Z_{22}i_{n+1}. \tag{I.2}$$

Our main problem will be to discuss the solutions of (I.2); in the process of discussing the transmission line, we shall find the properties of a four-terminal network as a special case of a line of one element.

2. Exponential Solution for Voltage and Current

The family of Eqs. (I.2), taken for all the values of n concerned in the transmission line, which we assume for the moment to be unlimited in both directions, can be solved by a simple assumption: we let both V_n and i_n vary exponentially with n. That is, we assume

$$V_n = V_0 e^{\gamma n}, \quad i_n = i_0 e^{\gamma n}, \qquad \text{(I.3)}$$

where V_0, i_0, γ, are constants. Substituting in (I.2), we find at once

$$V_0 = (Z_{11} + Z_{12} e^{\gamma}) i_0 = -(Z_{12} e^{-\gamma} + Z_{22}) i_0. \qquad \text{(I.4)}$$

Eliminating i_0 from (I.4), we have a quadratic for e^{γ},

$$e^{2\gamma} + \left(\frac{Z_{11}+Z_{22}}{Z_{12}}\right) e^{\gamma} + 1 = 0, \qquad \text{(I.5)}$$

whose solutions are

$$e^{\gamma} = -\frac{Z_{11}+Z_{22}}{2Z_{12}} \pm \left\{ \left(\frac{Z_{11}+Z_{22}}{2Z_{12}}\right)^2 - 1 \right\}^{\frac{1}{2}}. \qquad \text{(I.6)}$$

We readily find that the product of the two solutions of (I.6) is unity, so that the two are reciprocals of each other, and the corresponding values of γ are the negatives of each other. Remembering that all our quantities of course must be multiplied by the time variation $e^{j\omega t}$, we see that our two solutions, corresponding to the two signs for γ, represent traveling waves, the imaginary part of γ corresponding to a propagation constant, the real part to an attenuation constant; the two solutions represent waves propagated in opposite directions, so that one may be considered the direct wave, the other the reflected wave. If we write the complex number γ in terms of its real and imaginary parts,

$$\gamma = \alpha + j\beta, \qquad \text{(I.7)}$$

then we shall have two solutions, one with the real part α positive, the other with α negative. Henceforth we shall denote only that solution whose α is positive as γ, and shall explicitly call the other one $-\gamma$. Then we see at once that the solution involving $e^{-\gamma}$ represents the wave propagated to the right, that involving e^{γ} the wave propagated to the left, since the wave must be attenuated in the direction of propagation. The equations above determine not only γ; they also determine the ratio V_n/i_n for either of the solutions we have found. Finding V_n/i_n from (I.3) and (I.4), we have two alternative expressions; adding and dividing by two we get a symmetrical form,

$$\frac{V_n}{i_n} = \frac{Z_{11}-Z_{22}}{2} \pm Z_{12} \sinh \gamma = \zeta \mp Z_0, \qquad \text{(I.8)}$$

where we define

$$\zeta = \frac{Z_{11}-Z_{12}}{2}, \quad Z_0 = -Z_{12} \sinh \gamma. \qquad \text{(I.9)}$$

In these expressions, the upper sign goes with the solution e^{γ}, the lower sign with $e^{-\gamma}$.

3. The Terminated Line

In the preceding section, we have found two solutions of our problem:

$$V_n = (\zeta + Z_0) e^{-\gamma n}, \quad i_n = e^{-\gamma n}$$

and

$$V_n = (\zeta - Z_0) e^{\gamma n}, \quad i_n = e^{\gamma n}. \qquad \text{(I.10)}$$

In these expressions, we have seen that the first corresponds to a wave propagated toward the right, the second to a wave propagated toward the left. The general solution can be built up as a linear combination of these two. With an infinite line, any linear combination is a possible solution. On the other hand, if the line is only semi-infinite, consisting of sections stretching indefinitely to the left from the kth terminals, but is terminated with an impedance Z_k across those terminals, the wave traveling to the right can be treated as an incident wave, that traveling to the left as a reflected wave, and we find that the amplitude of the reflected wave is fixed if that of the incident wave is known. Let us assume a solution

$$V_n = A(\zeta + Z_0)e^{-\gamma n} + B(\zeta - Z_0)e^{\gamma n},$$
$$i_n = Ae^{-\gamma n} + Be^{\gamma n}, \qquad \text{(I.11)}$$

where A, B are coefficients to be determined. Their ratio can be found from the fact that $Z_k = V_k/i_k$. Thus we have

$$Z_k = \frac{\zeta + Z_0 + (\zeta - Z_0)(B/A)e^{2\gamma k}}{1 + (B/A)e^{2\gamma k}} = \zeta + Z_0 \frac{1 - (B/A)e^{2\gamma k}}{1 + (B/A)e^{2\gamma k}}. \qquad \text{(I.12)}$$

Solving (I.12), we have at once

$$\frac{B}{A}e^{2\gamma k} = \frac{Z_0 + \zeta - Z_k}{Z_0 - \zeta + Z_k}. \qquad \text{(I.13)}$$

4. Impedance of the Terminated Line

We now ask the question, what is the impedance V_n/i_n at the nth terminals, if a line is terminated by the impedance Z_k at the kth? To find the answer, we need merely take the ratio V_n/i_n from (I.11), substituting for B/A from (I.13). We find easily, denoting V_n/i_n by Z_n,

$$Z_n - \zeta = Z_0 \left(\frac{Z_0 \sinh \gamma(k-n) + (Z_k - \zeta) \cosh \gamma(k-n)}{Z_0 \cosh \gamma(k-n) + (Z_k - \zeta) \sinh \gamma(k-n)} \right). \qquad \text{(I.14)}$$

There are two interesting special cases: $Z_k - \zeta = 0$, and $Z_k - \zeta = \infty$. If $\zeta = 0$, a case which we often meet, the first corresponds to a short circuited line, and the second to an open circuited line. In these two cases we have

$$Z_n - \zeta = Z_0 \tanh \gamma(k-n),$$
or
$$Z_0 \coth \gamma(k-n), \qquad \text{(I.15)}$$

respectively. Equation (I.14) expresses the properties of the section of line between the nth and the kth terminals as a transformer: If an impedance Z_k is connected across the kth terminals, we find that the line transforms it into an impedance Z_n across the nth terminals. This relation can be expressed in another form which is sometimes more convenient, by introducing a quantity r_n by the definition

$$r_n = -\frac{B}{A}e^{2\gamma n} = \frac{Z_n - \zeta - Z_0}{Z_n - \zeta + Z_0}, \qquad \text{(I.16)}$$

where the second form is derived as is (I.12). The quantity r_n is the negative of the ratio of the reflected current to the incident current, at the nth terminals. Its value for $n = k$ is a measure of the reflection coefficient at the terminal impedance Z_k. It is convenient to denote r_n as the complex reflection coefficient, in general. We then have

$$\frac{r_n}{r_k} = e^{-2\gamma(k-n)},$$
$$\frac{Z_n - \zeta - Z_0}{Z_n - \zeta + Z_0} = e^{-2\gamma(k-n)} \left(\frac{Z_k - \zeta - Z_0}{Z_k - \zeta + Z_0} \right). \qquad \text{(I.17)}$$

Equation (I.17), which is equivalent to (I.14), is often more convenient, on account of its very simple dependence on n. It tells us that each section of the line which we go through multiplies the reflection coefficient by the complex factor $e^{-2\gamma}$. It may be used, for instance to give a simple solution of the following problem. In Section 1, we have characterized a single four-terminal network by three coefficients, Z_{11}, Z_{22}, and Z_{12}. From (I.6), (I.9), we can write these in terms of γ, ζ, and Z_0

$$Z_{11} = Z_0 \coth \gamma + \zeta, \quad Z_{22} = Z_0 \coth \gamma - \zeta,$$
$$Z_{12} = -Z_0 \operatorname{csch} \gamma. \qquad \text{(I.18)}$$

But now suppose we have, not a single network, but s identical networks connected together. They still form a four-terminal network, and can be characterized by new coefficients, which we may write as $(Z_{11})_s$, $(Z_{22})_s$, $(Z_{12})_s$, which, substituted into equations like (I.1), would give the relationship between the voltage and current at the input and output terminals of the composite network. It would be a considerable task to find these coefficients by straightforward manipulation. However, we note from (I.17) that the only difference between the transformation of the complex reflection coefficient produced by a single section of line, and by s sections, is that γ is multiplied by s in the second case. Thus, by retracing all our steps, this must be the only change introduced into (I.18), so that we have

$$(Z_{11})_s = Z_0 \coth \gamma s + \zeta,$$
$$(Z_{22})_s = Z_0 \coth \gamma s - \zeta, \qquad (I.19)$$
$$(Z_{12})_s = -Z_0 \operatorname{csch} \gamma s.$$

We may then write equations, similar to (I.2), for the relation between the currents in the nth, and the kth, terminals of our line

$$V_n = (Z_{11})_{n-k} i_n + (Z_{12})_{n-k} i_k,$$
$$V_k = -(Z_{12})_{n-k} i_n - (Z_{22})_{n-k} i_k, \qquad (I.20)$$

where the Z's are given in (I.19), substituting $n-k$ for s. If the kth terminals are connected by an impedance Z_k, we have $V_k = Z_k i_k$. Making that substitution, and solving for the impedance $Z_n = V_n/i_n$ seen at the nth terminals, we have

$$Z_n = (Z_{11})_{n-k} - \frac{(Z_{12})_{n-k}{}^2}{Z_k + (Z_{22})_{n-k}}$$
$$= \frac{(Z_{11})_{n-k} Z_k + (Z_{11})_{n-k}(Z_{22})_{n-k} - (Z_{12})_{n-k}{}^2}{Z_k + (Z_{22})_{n-k}}. \qquad (I.21)$$

The two forms of (I.21) are often used for expressing the transformer properties of a four-terminal network. They are equivalent to (I.14) and (I.17), as can be shown by straightforward manipulation.

5. Bilinear Transformations

In (I.14) and (I.21), we have two forms of the relation between the terminal impedance Z_k of a finite line of $n-k$ sections, and the transformed impedance which we see across the nth terminals. Both equations express Z_n as a bilinear function of Z_k; that is, as a function of the form

$$w = (az+b)/(cz+d), \qquad (I.22)$$

where the complex number w stands for Z_n, z for Z_k, and a, b, c, d are complex constants which have different meanings in the two cases. The properties of bilinear transformations are so important in our whole theory that we shall study them in some detail. We note in the first place that in any bilinear function like (I.22), we can divide numerator and denominator by c, so as to reduce the coefficient of z in the denominator to unity, as in (I.21). Thus such a function has three parameters, which can be uniquely determined from the impedance coefficients. We observe that (I.22) expresses a transformation of the complex number z into the complex number w. If we set up a complex plane for z, another for w, the transformation maps each point of the z plane onto the w plane, or *vice versa*. We can prove in a trivial manner that the inverse of a bilinear transformation (that is, the solution for z in terms of w) is again a bilinear transformation, so that the mapping of either plane onto the other is of the same nature. We can furthermore prove that the result of making two bilinear transformations in succession is itself a bilinear transformation. We now notice that since the function (I.22) is analytic, the mapping must be conformal, by fundamental principles of the theory of functions of a complex variable. That is, the shape of a small figure in the z plane is preserved in the corresponding figure in the w plane, though the scale in general will change, and the figure will be rotated. As a result of this conformal nature of the transformation, if two lines cross at a given angle in the z plane, the transformed lines will cross at the same angle in the w plane.

We next prove a property peculiar to the bilinear transformation: any circle in the z plane is mapped by the transformation into a circle in the w plane, and *vice versa*. To prove this, we note that there are two simple ways of expressing the equation of a circle. First, an equation

$$z\bar{z} + Az + \bar{A}\bar{z} + B\bar{B} = 0, \qquad (I.23)$$

where a bar indicates the complex conjugate, represents a circle. For if we let $z = R + jX$, this is

$$R^2 + X^2 + (A+\bar{A})R + j(A-\bar{A})X + B\bar{B} = 0, \qquad (I.24)$$

in which $A+\bar{A}$, $j(A-\bar{A})$, and $B\bar{B}$ are all real numbers, and this is obviously the equation of a circle in the z plane, in which R and X are coordinates. Secondly, an equation

$$w = C + \rho e^{j\phi}, \qquad (I.25)$$

where C is a complex constant, ρ a real number, ϕ a real number, represents a circle in the w plane, with center at C, radius ρ, provided ϕ takes on different values to represent parametri-

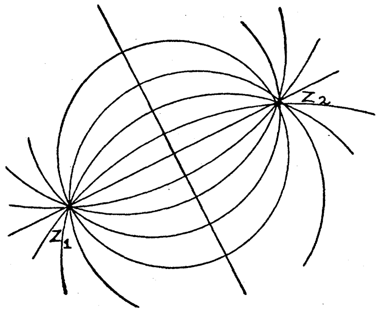

FIG. 3. Circles passing through characteristic impedances, in bilinear transformation.

cally the various points of the circle. Let us now represent w in this way, eliminate ϕ, and show that the resulting equation in z represents a circle. We have

$$\rho e^{i\phi}=\frac{az+b}{cz+d}-C,\quad \rho e^{-i\phi}=\frac{\bar{a}\bar{z}+\bar{b}}{\bar{c}\bar{z}+\bar{d}}-\bar{C}. \quad \text{(I.26)}$$

Multiplying these together, ϕ cancels out; rationalizing the denominator, we find at once that the equation for z is of the type described above, so that we have a circle in the z plane, resulting from transforming a circle in the w plane. As another mathematical theorem regarding the bilinear transformation, we shall prove that there are two particular values of z which are transformed into themselves, or remain invariant under the transformation. To find these, we need only require that $w=z$ in (I.22). Then we have $cz^2+(d-a)z-b=0$, a quadratic for z, whose solutions are

$$z=\frac{a-d}{2c}\pm\left\{\left(\frac{a-d}{2c}\right)^2+\frac{b}{c}\right\}^{\frac{1}{2}}. \quad \text{(I.27)}$$

The physical meaning of these values, in our case, is simple. They represent the impedances which must terminate the line, in order that the impedance across the nth terminals should be the same as across the kth terminals. That is, they must be just the values given by (I.8), characteristic of having only a direct or only a reflected wave. We find, in fact, that if we substitute the proper coefficients into (I.27) from either (I.14) or (I.21) we come out with just the values (I.8). These values are called the characteristic, or iterated, impedances of the line.

6. Graphical Discussion of Bilinear Transformations

The easiest way to visualize a bilinear transformation is to consider how certain lines in the z plane are transformed into corresponding lines in the w plane. Suppose we consider the two characteristic impedances, and the family of circles passing through them, in the z plane, as in Fig. 3. The characteristic impedances transform into themselves in the w plane, and any circle transforms into a circle. Thus each circle of this family must transform into another circle passing through the characteristic impedances, or into another circle of the same family. Consider similarly the family of circles orthogonal to these, in the z space. Each one of these must transform into another circle in the w space, which must by the conformal property of the transformation also be orthogonal to the family of circles passing through the characteristic impedances. In other words, each circle of this second family transforms into another circle of the same family. We can understand the exact nature of the transformation better if we consider the complex reflection coefficient, defined as in (I.16). That is, if z_1, z_2 are the two characteristic impedances, the two solutions of (I.27), we define

$$r=(z-z_1)/(z-z_2), \quad \text{(I.28)}$$

or, considering the actual impedances, we define

$$r_n=(Z_n-\zeta-Z_0)/(Z_n-\zeta+Z_0). \quad \text{(I.29)}$$

This is again a bilinear transformation. We note that when $z=z_1$, or $Z_n=\zeta+Z_0$, r or r_n is zero, and when $z=z_2$, or $Z_n=\zeta-Z_0$, r or r_n is infinite. The circles passing through z_1 and z_2 in the z plane then transform into straight lines through the origin in the r plane, and the orthogonal family of circles transform into concentric circles with the origin at the center. We may call such a transform of the impedance space a circle diagram. The bilinear transformation (I.14) or (I.21), when exhibited in the circle diagram, reduces to (I.17). That is, it corresponds to a multiplication of r_k by the constant factor $e^{-2\gamma s}$. The modulus of this factor corresponds to a change of scale along the radii, and the phase gives the rotation. Having found this interpretation of the bilinear transformation in the circle diagram, we can go back to the impedance plot, and see that the change of scale, which arises from the real part of γ, corresponds to a process of expansion of scale around one characteristic impedance, and a shrinking around the other, while the rotation corresponds to a process in which one of the circles surrounding one of the characteristic impedances transforms into itself.

7. Special Types of Networks and Lines

Let us now consider several special cases of networks and lines and their corresponding bilinear transformations, both in the impedance plane and the circle diagram. First we consider a lossless network, containing no resistances. In this case, a reactance connected across the terminals k must lead to a reactance seen across the terminals n. We can draw certain conclusions at once from this fact, using (I.21). If Z_k is infinite, or the network is open circuited, Z_n must be pure imaginary; thus $(Z_{11})_s$ must be pure imaginary. Similarly, open circuiting the terminals k, we find that $(Z_{22})_s$ must be imaginary. Short circuiting the terminals k, Z_n must be imaginary, or $(Z_{12})_s^2/(Z_{22})_s$ must be imaginary, from which $(Z_{12})_s^2$ must be real. An impedance Z_k with a positive resistive component must transform to a Z_n with a positive resistive component; thus $(Z_{12})_s^2$ must be negative, and $(Z_{12})_s$ is imaginary. Now we consider the transformation geometrically. Any point on the axis of ordinates in the Z_k plane must transform into another point on this axis in the Z_n plane, for all pure reactances are found on the axis of ordinates. There are only two ways in which the bilinear transformation can carry a straight line into itself. The first is that in which the straight line is the perpendicular bisector of the line joining the characteristic impedances, and in which the transformation in the circle diagram is a pure rotation, without change of scale. That is, the resistive parts of z_1 and z_2 are equal and opposite, so that ζ in (I.8) is pure imaginary, and Z_0 is real. Equation (I.9) then tells us that $\sinh\gamma$ is pure imaginary, or that γ is imaginary, equal to $j\beta$, and the factor $e^{-2\gamma s}$ becomes a pure rotation in the circle diagram. The other case arises when the two characteristic impedances are both located on the imaginary axis. In that case both ζ and Z_0 are pure imaginary, so that $\sinh\gamma$ and γ are real. In that case the transformation is the type which expands the scale around one characteristic impedance, contracts it around the other. The first of these two types of lossless networks is that found for instance in an ordinary section of wave guide or other transmission line, where the points in the circle diagram rotate on passing through a length of line. The second is that met in a wave guide beyond cut-off, in which all impedances in the Z_k plane tend to shrink toward the characteristic impedance in the Z_n plane.

The general case of a network with losses cannot be handled in such a simple way. The general bilinear transformation requires three complex constants to describe it, which we may take to be the values of the two characteristic impedances, and the value of γ. If the network is passive, that merely tells us that every value of Z_k in the right half plane (that is, every impedance with a real resistive component) must lie in the right half plane of the Z_n space. In general, however, it will transform into less than the whole half plane. The axis of ordinates in the Z_k plane must transform into a circle in the right half plane in the Z_n space, so that all physically allowable values of Z_k will transform into the interior of this circle in the Z_n space. We find, in general, that the greater the loss in the network, the smaller is the circle in the Z_n plane. Thus an attenuator in a line results in seeing almost the same impedance, regardless of what is attached to its other end.

8. Transformation of Resistance and Reactance Coordinates

In the general network with losses it is not particularly convenient to use the family of circles which we have so far considered, for the line of zero resistance does not form one of these circles. It is often convenient to visualize the transformation by drawing in the Z_n space the transforms of the lines $R_k=$constant, $X_k=$constant, or the rectangular coordinates in the k space. These form two orthogonal families of straight lines; they then transform into two orthogonal families of circles, as we see in Fig. 4. The circle $R_k=0$, bounding the region of positive resistances, forms a member of one of the families. The point $Z_k=\infty$ must lie on this circle, and each

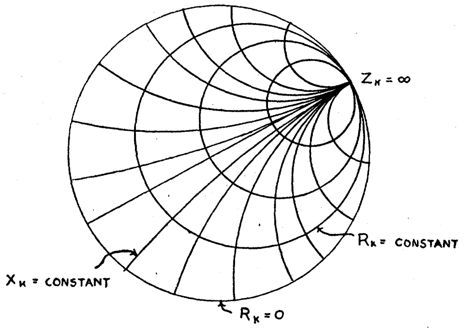

FIG. 4. Transforms of resistance and reactance coordinates, for general transformation.

of the circles X_k=constant must go through this point, and must cut the circle $R_k=0$ orthogonally. Each circle R_k=constant must also go through this point, and must be tangent to the circle $R_k=0$ there. Thus the plot of lines of constant R_k and constant X_k, in the Z_n plane, must have the form shown in Fig. 4. This, it must be emphasized, is perfectly general, holding for an arbitrary network with losses. The three complex constants characterizing the network, in this way of exhibiting it, may be taken to be the coordinate of the center of the circle; the radius, and orientation of the point of infinite impedance Z_k (the modulus and phase of a complex number); and the location, within the circle, of that characteristic impedance which lies within.

9. The Continuous Transmission Line

A continuous transmission line, such as a parallel wire line or a wave guide, can be considered as the limiting case of the type of line we have been considering, as the separate elements become smaller and smaller, and closer and closer together.* We may get the relations for this continuous line easily by passing to the limit from the various formulas which we have developed. In such a continuous line, the voltage and current are given as functions of distance along the line by formulas like (I.11), in which we need make only two modifications: first, such lines are symmetrical, one end being equivalent to the other, which means that ζ is zero; secondly, we must let n refer to the distance measured along the line, which would be the case if we regarded each unit length of the line as one of the networks of our previous discussion. In the reference quoted above, it is shown how the characteristic impedance Z_0 and the propagation constant γ can be found from distributed constants of the line, by equations which represent the passage to the limit of (I.6) and (I.9), but we shall not need to use those equations. With the small changes noted, the whole discussion of Sections 4–7 applies to continuous lines. Thus, in particular, we can introduce a reflection coefficient, by (I.16), equal to $(Z-Z_0)/(Z+Z_0)$, where Z is the impedance at an arbitrary point of the line, and as in (I.17), the reflection coefficient is multiplied by a factor $e^{-2\gamma s}$ if we travel backward a distance s along the line. We can regard a length s of transmission line as a transformer, with transformer coefficients given by (I.19), setting $\zeta=0$.

We are often interested in the lossless line. In that case, as we saw in Section 7, there are two possible cases. First, we can have Z_0 real, γ pure imaginary. In that case there is real propagation along the line, with no attenuation, and the effect of traveling along the line a certain distance is to rotate through a certain angle in the circle diagram, a rotation of 360° corresponding to a half-wave-length. The rotation is positive, or counter-clockwise, as we go away from the generator toward the load, or negative as we go toward the generator. In the impedance plane, we may set up circles like those in Fig. 5. The circles surrounding the characteristic impedances represent circles on which the magnitude of the reflection coefficient r is constant, and the circles passing through the characteristic impedances are circles of constant phase of reflection coefficient. Traveling along the line toward the generator, we travel in a negative direction about the first family of circles. Once in a half-wave-length the impedance goes through a maximum real value (we shall presently show that this corresponds to the standing wave maximum), and a quarter-wave-length further along it goes through a minimum (the standing wave minimum). The axis of ordinates forms a limiting case of these circles, corresponding to the case where the terminal impedance is purely reactive. In that case, the maximum impedance is infinite,

* For a discussion of this limit, see J. C. Slater, *Microwave Transmission* (McGraw-Hill Book Company, Inc., New York, 1942), pp. 17–21.

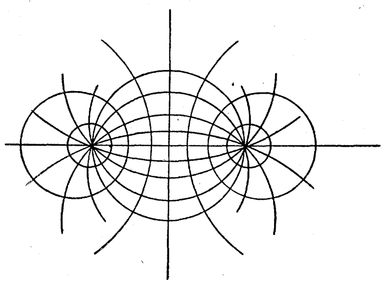

FIG. 5. Circles of constant reflection coefficient and phase, in the impedance plane.

the minimum is zero. All real loads lie in the right half plane, corresponding to positive resistive components. Correspondingly, in the circle diagram, as in Fig. 6, all real loads correspond to reflection coefficients within a unit circle, the circle into which the axis of ordinates in the impedance plane transforms. To see this, we need only notice that if we set $Z=0$ or $Z=\infty$ in the expression $r=(Z-Z_0)/(Z+Z_0)$, we find $r=-1$ and $+1$, respectively, so that these are two points on the circle representing a purely reactive load. Corresponding to Fig. 4, we can draw lines of constant resistance and constant reactance in the circle diagram. Since we have just seen that the point $Z=\infty$ transforms to the point $r=1$, the circles of constant X all pass through this point, and the circles of constant R are tangent to unit circle there. This diagram can be used as a very convenient way for finding the transformation produced by a length of lossless line on a given terminal impedance. If the lines of constant R and X are properly labeled, we can at once look up on the diagram the point corresponding to a given terminal impedance. If we choose, we can find the modulus and phase of the reflection coefficient, by measuring the radius and phase angle of the radius vector out to the corresponding point, from the center of the circle. To travel a given distance along the line, we then merely rotate the radius vector through an appropriate angle, and read off the transformed values of R and X. This construction is made the basis of a convenient rotating slide rule for calculating impedances.

The other case of a lossless transmission line is that in which Z_0 is imaginary, γ real. This corresponds, for instance, to a wave guide beyond cut-off. There is only attenuation, no propagation, down the line, and the effect of traveling along the line is to shrink the scale in the circle diagram, without rotation, so that after traveling such a distance down such a line, no matter what the terminal impedance may be, the impedance seen looking into the guide is very close to the characteristic impedance, which we remember is purely imaginary. In Fig. 7, showing the impedance plane, the effect of passing down the line away from the load is to travel along the circles passing through the characteristic impedance, as shown by the arrows. In the circle diagram, the role of R and X is interchanged with respect to the case of propagation on the lossless line. The horizontal axis becomes the line $R=0$, the positive resistances lying in the lower half plane, and the unit circle becomes the line $X=0$, so that we need the area outside as well as inside the unit circle to describe all real terminal impedances. We can again use the circle diagram to represent the effect of a length of line: traveling down the line a certain distance away from the load shrinks the radius vector by an appropriate factor, without rotation, so that as we have mentioned before all impedances approach closer and closer to the characteristic impedance as the length of the attenuating line becomes greater.

10. Standing Wave Ratios

The case of a lossless continuous line is a very important one practically, for wave guides are used in practice for measuring standing wave ratios, and hence terminal impedances. We, therefore, next consider the definition of standing wave ratios in a continuous lossless line in which

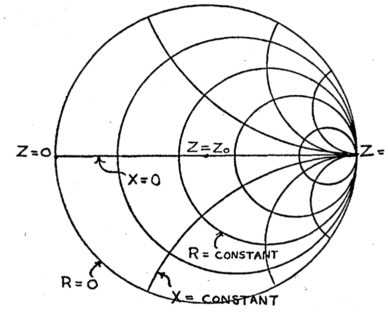

FIG. 6. Circles of constant resistance and reactance, in the reflection coefficient plane.

there is real propagation. The measurement of standing wave ratio involves a measurement of voltage as a function of distance along the line. What essentially is done is to put a very high impedance shunt across the line, and measure the power generated in that shunt; the power will be the square of the modulus of the voltage, multiplied by the conductance of the shunt. From (I.11), the voltage as a function of distance n along the line will be

$$V_n = Z_0(Ae^{-\gamma n} - Be^{\gamma n}) = Z_0(Ae^{-i\beta n} - Be^{i\beta n}). \quad \text{(I.30)}$$

Introducing the reflection coefficient

$$r = -(B/A)e^{2i\beta n} \quad \text{(I.31)}$$

from (I.16), we see that the voltage can be written

$$V_n = Z_0 A e^{-i\beta n}(1+r), \quad \text{(I.32)}$$

so that the power delivered to the shunt is proportional to

$$|V_n|^2 = Z_0^2 |A|^2 (1+r)^2 \quad \text{(I.33)}$$

and the magnitude of the voltage is proportional to $|(1+r)|$. In Fig. 8 we see the locus of the points $1+r$ as we go along the transmission line. It is clearly a circle of radius $|r|$, with center at the point 1. The radius vector from the origin out to this circle then has a length proportional to the voltage, so that the square of its length is proportional to the power delivered to the shunt, as in (I.33). It is now clear that the maximum length of this vector comes when r is real and positive, when its value is $1+|r|$, so that these

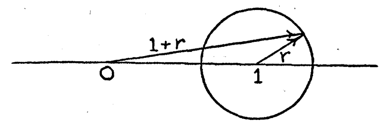

FIG. 8. Diagram for interpreting standing wave ratio.

points, as we have mentioned earlier, are the standing wave maxima, or points at which maximum power is delivered to the standing wave detector; and the minimum length comes when r is real and negative, when its value is $1-|r|$, and these points are the standing wave minima. We now define the standing wave ratio in voltage as the ratio of maximum to minimum voltage

$$SWR(\text{voltage}) = \frac{1+|r|}{1-|r|}. \quad \text{(I.34)}$$

We furthermore define the standing wave ratio in power as the ratio of maximum to minimum power

$$SWR(\text{power}) = \left(\frac{1+|r|}{1-|r|}\right)^2. \quad \text{(I.35)}$$

It is often convenient also to define a standing wave ratio in decibels, as the number of decibels by which the power delivered to the standing wave detector at standing wave maximum must be attenuated to make it equal to the power delivered at standing wave minimum. We have

$$SWR(db) = 10 \log_{10} SWR(\text{power}) = 20 \log_{10} SWR(\text{voltage}). \quad \text{(I.36)}$$

It is now clear that by a measurement of the standing wave ratio we can find the magnitude of the reflection coefficient, $|r|$. Also from the position of the standing wave minima and maxima along the line, we can find the phase of the reflection coefficient: in the circle diagram, r has a phase angle of zero at standing wave maxima, and of 180° at standing wave minima. Thus a measurement of standing wave ratio and position of standing wave minimum allows us to find the impedance seen across any arbitrary plane of the continuous transmission line. A given standing wave ratio corresponds to a given radius in the circle diagram, as we see from (I.34), and the phase angle of rotation corresponding to a given plane is simply found by measuring the distance from standing wave maximum to the given plane, and noting that a half-wave-length rotates by

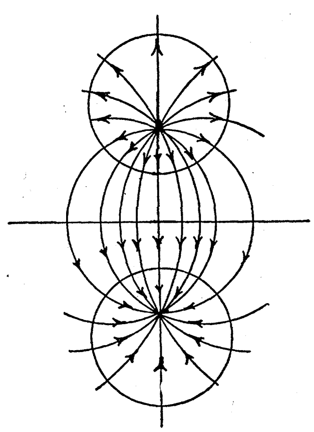

FIG. 7. Transformation produced by lossless line beyond cut-off.

360°. It is interesting to note that the impedance at standing wave maxima and minima is easily found from the standing wave ratio. Solving the equation $r=(Z-Z_0)/(Z+Z_0)$ for Z/Z_0, we have

$$(Z/Z_0)=(1+r)/(1-r). \tag{I.37}$$

At standing wave maximum, $r=|r|$; thus, comparing with (I.34), we see that Z/Z_0 at standing wave maximum equals the standing wave ratio in voltage. At standing wave minimum, $r=-|r|$; thus Z/Z_0 at standing wave minimum equals the reciprocal of the standing wave ratio in voltage. We note, then, that the characteristic impedance Z_0 is the mean proportional between the values of impedance at standing wave minimum and maximum.

11. Transformers between Transmission Lines

In many microwave applications, two sections of wave guide are connected together by a lossless section of some sort, acting as a transformer between the two guides. It can be merely an iris or a change of cross section, or a much more complicated object, such as a resonant cavity with two separate wave guide outlets. The corresponding circuit is a four-terminal network, without losses, with a continuous transmission line connected to each of its pairs of terminals, as shown in Fig. 9. We can get interesting and valuable results by considering the impedance transformation between points 1, 2, arbitrary points of the two guides. The whole circuit between 1 and 2 is, of course, itself a four-terminal network, which can be handled by the same methods we have already employed. We shall, however, write the impedance transformation in a different form from what we have used before. In the first place, we shall make a transformation, not of the impedance, but of the reflection coefficient. We shall set this up, however, in a way which differs essentially from that used, for instance, in Eq. (I.17). In that case, we defined the reflection coefficient with respect to the characteristic impedances of the whole network between terminals 1 and 2. Here, on the contrary, we shall define r_1 as the reflection coefficient with respect to the characteristic impedance of the transmission line to the left of the network, and r_2 as the reflection coefficient with respect to the characteristic impedance of the line to the right. That is, if Z_{01}, Z_{02} are the characteristic impedances of these two lines, we have

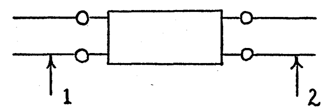

FIG. 9. Four-terminal network with transmission lines.

$$r_1=\frac{Z_1-Z_{01}}{Z_1+Z_{01}}, \quad r_2=\frac{Z_2-Z_{02}}{Z_2+Z_{02}}, \tag{I.38}$$

where Z_1, Z_2 are the impedances at the two points. We now know that there must be a bilinear transformation connecting Z_1 and Z_2. Since r_1 and Z_1, and r_2 and Z_2, are connected by the bilinear transformation (I.38), we know that r_1 and r_2 are connected by a bilinear transformation. Furthermore, the transformer is lossless; thus a point on the unit circle in the r_2 space must transform into a point on unit circle in the r_1 space, since the unit circle in each case represents a pure reactance. A lossless transformer is characterized by three parameters, as we saw for instance in Section 7, where for the lossless case the three parameters Z_{11}, Z_{12}, and Z_{22}, which in the general case are complex and hence amount to six independent numbers, are pure imaginary. We shall now set up a bilinear transformation between r_1 and r_2 which satisfies all these conditions, and which thus is a general expression of the transformation. This is

$$r_1=e^{i\phi_1}\left(\frac{D+r_2e^{-i\phi_2}}{1+Dr_2e^{-i\phi_2}}\right), \tag{I.39}$$

where ϕ_1, ϕ_2, and D are three real constants. Let us check that this transformation has the required properties. First, it is clearly a bilinear transformation, with three independent parameters. Next we must show that it transforms unit circle into unit circle. If r_2 is a point on unit circle (that is, if its magnitude is unity), then clearly $r_2e^{-i\phi_2}$, which simply amounts to the same vector rotated through angle $-\phi_2$, is also on unit circle; similarly if r_1 is on unit circle, $r_1e^{-i\phi_1}$ is on unit circle. We must then merely prove that if r_2 is on unit circle, $(D+r_2)/(1+Dr_2)$ is also on unit circle. We note first that the points ± 1 are invariant under this transformation: if $r_2 = \pm 1$, $r_1 = \pm 1$ also. Thus the transform of unit circle in the r_2 space, into the r_1 space, passes through these two points. Furthermore, the transformed circle in the r_1 space is symmetrical about the axis of abscissae. We can see this by noting that the transforms of the points $\pm j$ in the r_2 space are $(D \pm j)/(1 \pm jD)$ in the r_1 space, which are complex conjugates of each other, and hence are located symmetrically with respect to the real axis. This fixes the transform of unit circle into the r_1 space uniquely, and shows that it is again unit circle. The transformation $(D+r_2)/(1+Dr_2)$ is in fact the sort in which there is a motion along those circles passing through the fixed points of the transformation (which in this case are ± 1), without rotation, as shown in Fig. 10; such a transformation must always carry the circle passing through the two fixed points, and with its center at the point midway between the two fixed points, into itself. Now that we have justified the form (I.39) for the transformation, we can give it a very interesting physical interpretation. In the first place, $r_1 e^{-j\phi_1}$ represents the reflection coefficient, not across the point 1, but across another point of the same transmission line; for we remember that the reflection coefficient is multiplied by the factor $e^{-2j\beta s}$ in moving a distance s toward the generator. We may call the reflection coefficient across this point r_1', which is equal to $r_1 e^{-j\phi_1}$. Similarly $r_2 e^{-j\phi_2}$ is the reflection coefficient r_2' across another point in the other transmission line. The transformation between these quantities is then

$$r_1' = (D + r_2')/(1 + Dr_2'). \qquad \text{(I.40)}$$

We can easily prove from this that

$$\frac{1+r_1'}{1-r_1'} = \frac{1+D}{1-D}\,\frac{1+r_2'}{1-r_2'}. \qquad \text{(I.41)}$$

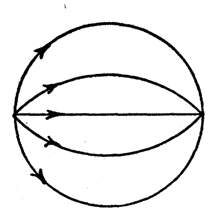

FIG. 10. Multiplication of impedance by real constant, in reflection coefficient plane.

But by (I.37), this is equivalent to

$$\frac{Z_1'}{Z_{01}} = \frac{1+D}{1-D}\,\frac{Z_2'}{Z_{02}}, \text{ or } Z_1' = \left(\frac{Z_{01}}{Z_{02}}\,\frac{1+D}{1-D}\right) Z_2'. \qquad \text{(I.42)}$$

In other words, between the points 1′ and 2′, the transformation is one of multiplying the impedance by a real constant. We have then the valuable general theorem: if we have any type of lossless network, with two outputs in the form of transmission lines, we can find sets of points on each of the lines, such that the impedance across one of the points is a real factor times the impedance across the other.

12. Determination of Transformer Constants

It is often very important to determine experimentally the three constants characterizing a lossless transformer connecting two transmission lines, such as we have just been discussing. There are several ways to do this. For example, we may place the characteristic impedance Z_{02} across the terminals 2, or match the right-hand line. Then r_2 will be zero, and we shall have $r_1 = De^{j\phi_1}$, so that measuring the standing wave ratio and position of standing wave minimum in the left-hand line, we measure D and ϕ_1. Similarly we can match the left-hand line and measure the standing wave ratio and position of standing wave minimum in the right-hand line. We must be careful how we apply the formulas to this case, for we have set them up only for the case where the power is flowing to the right; but it turns out, as we should expect, that from this measurement we can determine D and ϕ_2. An experimental check is to see whether the values of D obtained in the two cases agree. This method of finding the constants, though sometimes convenient, is not ideal, in that both transmission lines have to be equipped so that we can insert matched loads into them, and so that we can measure standing waves in them. Often a more convenient measurement is that of inserting a short circuit (that is, using wave guides, a plunger) at an adjustable point along the right-hand transmission line, and measuring the position of standing wave minimum in the left-hand line as a function of the plunger position. Let us analyze this experiment. We assume the short circuit to be located a distance d_2 to the right of the point 2. Then we shall have

$$r_2 = -e^{-2j\beta_2 d_2}, \qquad (I.43)$$

where β_2 is the value appropriate to the right-hand line. In this case we can show easily that

$$\frac{Z_2'}{Z_{02}} = j \tan\left(\frac{\phi_2}{2} + \beta_2 d_2\right). \qquad (I.44)$$

In the left-hand line, there will be an infinite standing wave ratio, or a reflection coefficient of unit magnitude, since we shall be seeing a purely reactive impedance. Thus at a standing wave minimum the impedance will be zero, or there will be an effective short circuit. Let a standing wave minimum be located a distance d_1 to the right of the point 1. Then by an argument just like that used above we have

$$\frac{Z_1'}{Z_{01}} = j \tan\left(\frac{\phi_1}{2} + \beta_1 d_1\right). \qquad (I.45)$$

We then have, as a result of (I.42),

$$\tan\left(\frac{\phi_1}{2} + \beta_1 d_1\right) = \frac{1+D}{1-D} \tan\left(\frac{\phi_2}{2} + \beta_2 d_2\right). \qquad (I.46)$$

This equation is in a convenient form for practical use. If we plot $\beta_1 d_1$ as a function of $\beta_2 d_2$, we get a curve of the form shown in Fig. 11, sometimes called an S curve on account of its shape. The coordinates of the points of maximum and minimum slope determine ϕ_1 and ϕ_2, as shown, and the slope at these points determines D: the maximum slope is $(1+D)/(1-D)$, and the minimum is its reciprocal.

13. A Lossless Transformer as a Shunt or Series Reactance

In Section 11 we have shown that a lossless transformer between two transmission lines can be described as an impedance multiplication by a real transformer ratio between certain definite points in the two lines. There is another equally legitimate way of considering an arbitrary lossless transformer, which is sometimes useful. Using the same diagram of a transformer that we have given in Fig. 9, we may write the transformer equations in either of the forms

$$\frac{1 - r_1 e^{-j\phi_1}}{1 + r_1 e^{-j\phi_1}} = \frac{1 - r_2 e^{-j\phi_2}}{1 + r_2 e^{-j\phi_2}} + jy, \qquad (I.47)$$

$$\frac{1 + r_1 e^{-j\phi_1}}{1 - r_1 e^{-j\phi_1}} = \frac{1 + r_2 e^{-j\phi_2}}{1 - r_2 e^{-j\phi_2}} + jx. \qquad (I.48)$$

These relations are both bilinear transformations (we can see this easily if we solve them for r_1 in terms of r_2); they have three arbitrary constants, if ϕ_1, ϕ_2, and y or x are real. Furthermore, they transform reactances into reactances. To show this, we let the magnitude of r_1 be unity. That is, we let $r_2 = e^{-j\theta_2}$, where θ_2 is real. Then the expression on the right of (I.47) becomes $j \tan(\theta_2 + \phi_2)/2$, which is pure imaginary. Adding jy, we still have a pure imaginary. Reversing the argument, this shows that r_1 has unit magnitude, or that the impedance Z_1 is a pure reactance. The meaning of (I.47) or (I.48) is simple. The quantity on the left of (I.48) is the ratio of the impedance, to the characteristic impedance, at a certain point on the line, as we see from (I.37). The quantity on the right has a similar interpretation. Thus we see that there is a certain point on the right-hand line, and a corresponding point on the left-hand line, such that the impedance seen at the point on the left line is a series combination of the impedance at the point on the right, and a fixed series reactance jx. Similarly the quantities in (I.47), being the reciprocals of the impedances, are admittances: there are corresponding points on the two lines such that the admittance across the left-hand point is the shunt combination of the admittance across the right-hand point, and a fixed shunt admittance jy. It is to be understood that the phase angles ϕ_1 and ϕ_2 are in general different in (I.47) and (I.48), and in turn different from those in (I.39); that is, there are different sets of corresponding points in the two lines, with respect to which the transformer is an ideal transformer (that is, the impedance is multiplied by a given ratio, as in (I.41)); or a shunt reactance (as in (I.47)) or a series reactance (as in (I.48)). It would be possible to get the relations between the various sets of constants, but it is not very profitable to do so. The expression of a lossless transformer as a shunt reactance is particularly

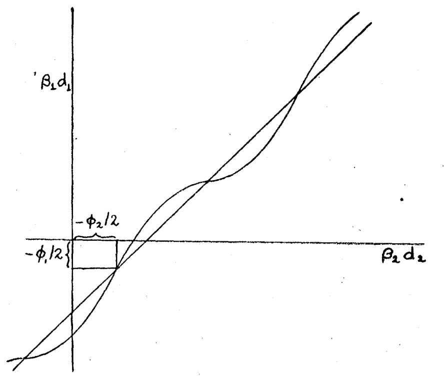

FIG. 11. An S curve.

appropriate in such cases as the transformer produced by an iris in a diaphragm in a wave guide. In such a case, provided the hole in the iris is small, we can show that, if points 1 and 2 are immediately on the two sides of the diaphragm, the angles ϕ_1 and ϕ_2 are practically zero. In other words, the diaphragm acts just like a shunt reactance, without any correction terms. In more general cases, however, a transformer will act like a combination of a shunt reactance, plus an effective lengthening or shortening of the line on each side. It is interesting to realize, as a result of the discussion of the last three sections, that any desired impedance transformation can be secured in many ways, including shunt and series reactances plus appropriate change of length of lines. For instance, suppose we have a change of characteristic impedance between two guides. We can always insert an iris in such a way as to match the guides into each other, in the sense that a matched load ($r_2=0$) in one guide transforms into a matched load ($r_1=0$) in the other.

14. Power Flow in Networks and Transmission Lines

So far, our discussion has dealt almost entirely with the impedance, the ratio of voltage to current. Of almost equal importance is the power, or product of voltage and current. From the two together, we can work backward and find voltage and current separately. We shall find, in our microwave work, that it is more convenient to deal with impedance and with power flow than with voltage and current; for impedance and power flow are more readily measurable. Let us first find the formula for power flow into a pair of terminals between which there is a voltage V, and a current i, the sign convention being the one we have been continually using. We must remember that when we write V as the voltage, we really mean that the voltage is the real part of $Ve^{j\omega t}$. That is, if V is written in terms of its real and imaginary parts, as V_r+jV_i, the voltage is $V_r\cos\omega t-V_i\sin\omega t$. Similarly the current is $i_r\cos\omega t-i_i\sin\omega t$, writing in a corresponding way. The power flow is the product of voltage and current; that is, it is

$$\text{Power}=V_r i_r\cos^2\omega t+V_i i_i\sin^2\omega t-(V_r i_i+V_i i_r)\sin\omega t\cos\omega t. \quad (\text{I}.49)$$

This is the instantaneous power, which as we see involves squares and products of $\sin\omega t$ and $\cos\omega t$. Ordinarily we are interested only in the time average of the power. Averaging, the average of $\cos^2\omega t$ and of $\sin^2\omega t$ is $\frac{1}{2}$, and of $\sin\omega t\cos\omega t$ is zero. Thus if P denotes the average power, we have

$$P=\tfrac{1}{2}(V_r i_r+V_i i_i). \quad (\text{I}.50)$$

This can, however, be written in another way. If we take the product $V\bar{i}$, we have

$$(V_r+jV_i)(i_r-ji_i)=(V_r i_r+V_i i_i+j(V_i i_r-V_r i_i)).$$

The real part of this quantity leads to the power as in (I.50). Thus we have

$$P=\tfrac{1}{2}Re\,V\bar{i}. \quad (\text{I}.51)$$

This is the standard formula for average power, in terms of complex amplitudes of voltage and current. If $V=Zi=(R+jX)i$, we then have

$$P=\tfrac{1}{2}|i|^2R, \quad (\text{I}.52)$$

where $|i|$ is the magnitude of the current.

The formula (I.51) or (I.52) can be applied at once to the flow of power in a transmission line. Let us assume the solutions (I.11) for voltage and current at the nth terminals of a line, taken together with the definition (I.16) for the complex reflection coefficient. We may then write voltage and current in the form

$$V_n=\zeta i_n+Z_0(Ae^{-\gamma n}-Be^{\gamma n})=\zeta i_n+Z_0Ae^{-\gamma n}(1+r_n),$$
$$i_n=Ae^{-\gamma n}+Be^{\gamma n}=Ae^{-\gamma n}(1-r_n). \quad (\text{I}.53)$$

If we let $Z_0=R_0+jX_0$, we then have

$$P=\tfrac{1}{2}|i_n|^2 Re\zeta+\tfrac{1}{2}R_0(A\bar{A}e^{-2\alpha n}-B\bar{B}e^{2\alpha n})$$
$$+\tfrac{1}{2}X_0 j(A\bar{B}e^{-2j\beta n}-\bar{A}Be^{2j\beta n}),$$
$$=\tfrac{1}{2}|i_n|^2 Re\zeta+\tfrac{1}{2}R_0 A\bar{A}e^{-2\alpha n}(1-r_n\bar{r}_n)$$
$$+\tfrac{1}{2}X_0 A\bar{A}e^{-2\alpha n}j(r_n-\bar{r}_n), \quad \text{(I.54)}$$

where $\gamma=\alpha+j\beta$, as before. Let us see what (I.54) becomes in certain simple cases. In a network without losses, which is propagating the wave, we have seen that ζ is imaginary, and Z_0 is real, so that $X_0=0$. Furthermore $\alpha=0$. Thus in this case we have

$$P=\tfrac{1}{2}Z_0 A\bar{A}(1-r_n\bar{r}_n)=\tfrac{1}{2}Z_0(A\bar{A}-B\bar{B}). \quad \text{(I.55)}$$

Thus in this case the power flow is simply the difference between the flows of power in the direct and reflected waves. The quantity $r_n\bar{r}_n$ is independent of n; for if we let

$$r_n=\rho e^{-2j\beta n}, \quad \text{(I.56)}$$

where ρ is the magnitude of the reflection coefficient, we see that $r_n\bar{r}_n=\rho^2$. On the other hand, in a network without losses in which there is attenuation, we have seen that ζ is again imaginary, Z_0 is imaginary, and γ is real, equal to α. Thus in this case we have

$$P=\tfrac{1}{2}X_0 j(A\bar{B}-\bar{A}B). \quad \text{(I.57)}$$

The power flow is again independent of n, but in this case there is no flow unless A and B are both different from zero. In other words, a purely attenuated wave, in for instance a wave guide beyond cut-off, carries no power. The reason for this is simple. At one end of the guide, at infinite distance, the amplitude is attenuated to zero, and obviously no power can be flowing there. But since there are no losses along the guide, all the power that flows in one end must flow out the other. Thus there can be no power flow anywhere. On the other hand, if the guide is only of finite length, both A and B must be different from zero, to satisfy boundary conditions, as indicated for instance by Eq. (I.13). Power will then flow; and this is reasonable physically, since a certain amount can pass through such a guide beyond cut-off into a terminal impedance. The problem mathematically and physically is like that met in total internal reflection in optics. The totally internally reflected wave results in an exponentially damped wave in the rare medium. If that medium extends to infinity, the exponential wave carries no energy, and all the energy is reflected. If, on the other hand, there is only a thin sheet of rare medium, then another dense medium, we shall have to have both sorts of exponential waves to satisfy the boundary conditions, and we shall find that power is transmitted through the rare medium to the adjacent dense medium.

Another interesting case is the guide or other transmission line with a real characteristic impedance, or $X_0=0$, but with a slight attenuation, so that α is different from zero. If $\zeta=0$, as it is in a guide, then we have

$$P=\tfrac{1}{2}Z_0(A\bar{A}e^{-2\alpha n}-B\bar{B}e^{2\alpha n}). \quad \text{(I.58)}$$

This has a simple interpretation: the first term is the power flow in the direct wave, which varies as $e^{-2\alpha n}$ on account of the decrease of amplitude of this wave as n increases; the other term is the flow backward in the reflected wave, which decreases as n decreases. In this case the magnitude ρ of the reflection coefficient depends on n.

15. Power Flow from a Lossless Line into a Terminal Impedance

If we have power flowing from a lossless transmission line into a terminal impedance, we have seen in (I.55) that the power flow is $\tfrac{1}{2}Z_0 A\bar{A}(1-r\bar{r})$. By (I.38), $r=(Z-Z_0)/(Z+Z_0)$. If $Z=R+jX$, we find easily that

$$\frac{P}{\tfrac{1}{2}Z_0 A\bar{A}}=1-\frac{(R-Z_0+jX)(R-Z_0-jX)}{(R+Z_0+jX)(R+Z_0-jX)}$$
$$=\frac{4RZ_0}{(R+Z_0)^2+X^2}. \quad \text{(I.59)}$$

The expression (I.59) tells how the power absorbed in the terminal impedance varies with R and X, provided $A\bar{A}$ remains constant. This means that the power flow in the direct wave is constant, independent of the power flow in the reflected wave (for of course that changes, as the load changes). We can secure this situation in practice by taking a generator of power, then inserting an attenuator of high attenuation, then our transmission line and load. This is called padding the generator with an attenuator. The object of the attenuator is to absorb practically all of the reflected wave before it gets back to the power source, so that the reflected wave will not react back on the source, and change its power output. Put another way, we have found that if there is an attenuating section in a line, the input impedance seen looking into the attenuating section is almost independent of the output impedance. That is, the impedance seen by the generator is almost independent of the load impedance; so much of the power is absorbed by the attenuator that the small remaining power absorbed by the load is negligible. Since the power output of a generator depends on the load which it sees, this means that the power output is practically constant as Z is varied, so that the incident amplitude A is practically constant. We must note at the same time, however, that A represents the amplitude of the power delivered by the generator through a high attenuation, so that this procedure is very wasteful of power. When we have this condition, which is often used for measurement, we then have an output power proportional to (I.59). It is interesting to consider this function as it depends on R and X. Obviously as far as X is concerned, the power is a maximum when $X=0$, or when the load is purely resistive. If furthermore we vary R to make the power a maximum, we find by differentiating with respect to R and setting the derivative equal to zero that the maximum comes when $R=Z_0$, or when the load impedance equals the characteristic impedance of the line, or is a matched load; in this case the function equals unity. It is interesting to consider the contours of constant power in a Z plane, in which R and X are plotted as variables. Since the power depends on the magnitude of the reflection coefficient, the power will be constant on a circle of constant magnitude of reflection coefficient, or constant ρ. In Fig. 5 we showed a diagram of these circles, a family of circles surrounding the characteristic impedances of the line. The function (I.59) is unity at the right-band characteristic impedance, and decreases as we go away from that point, becoming zero along the imaginary axis, or for a reactive load, which of course can absorb no power. In the left-hand half plane, the function is negative; such loads have negative resistive components, and correspond to generators, rather than passive loads, so that in the presence of such a terminal impedance the power would flow to the left, or (I.59) should be negative. The function is negatively infinite at the left-hand characteristic impedance; the meaning of this is that at that characteristic impedance the wave is flowing entirely to the left, so that the wave traveling to the left is infinite if A, the amplitude of the wave traveling to the right, is finite. These features of the left-hand half plane are not of physical interest in ordinary applications. In the reflection coefficient plane, the contours of constant power are of course circles concentric with the origin, the power being a maximum at the origin, going down to zero at unit circle, and to a negatively infinite value at infinity.

16. Circuit Efficiency and Insertion Loss of a Resistive Network

If we have a network of the general type characterized by transformer coefficients Z_{11}, Z_{22}, Z_{12} in the manner of Eq. (I.2), and terminated by an impedance Z, we shall find that some of the power fed into the left-hand terminals of the network (which we shall call terminals 1) is absorbed in the network, and some in the impedance Z. If we let P_1 be the power flowing across the terminals 1, and p_2 the power across the terminals 2 (the terminals connecting the right-hand side of the network to Z), then the fraction P_2/P_1 of the input power will be delivered to the load. If the object of the network is to deliver power, as it is in some practical cases, we may call this ratio the circuit efficiency, and denote it by η_c. Thus we have

$$\eta_c=\frac{P_2}{P_1}=\frac{Re(V_2\bar{\imath}_2)}{Re(V_1\bar{\imath}_1)}. \qquad \text{(I.60)}$$

The circuit efficiency will always be less than unity if there are losses in the network; it will be zero if the output load is reactive, and can absorb no power; and it is a measure of the effectiveness of the network as a carrier of power. Sometimes we are interested in a network as an attenuator, and in that case we are interested in the amount by which it decreases the power passing through it. In that case we define an insertion loss, a measure of the decrease of power in passing through the network, measured in decibels. That is, we have

$$\text{Insertion loss} = 10 \log_{10} P_1/P_2 = 10 \log_{10} (1/\eta_c). \quad (I.61)$$

We shall now calculate the circuit efficiency as a function of the terminal impedance Z; from it the insertion loss can be found from (I.61).

From (I.2) and (I.21) we have

$$V_1 = \left(Z_{11} - \frac{Z_{12}^2}{Z+Z_{22}}\right) i_1, \quad \frac{i_2}{i_1} = \frac{-Z_{12}}{Z+Z_{22}}. \quad (I.62)$$

Writing the output impedance as $Z = R + jX$, and writing $Z_{11} = R_{11} + jX_{11}$, etc., and remembering that $V_2 = i_2 Z$, we have

$$\eta_c = \frac{R|i_2|^2}{Re\left(Z_{11} - \dfrac{Z_{12}^2}{Z+Z_{22}}\right)|i_1|^2}. \quad (I.63)$$

Taking the values of the complex quantities, we have

$$Re\left(Z_{11} - \frac{Z_{12}^2}{Z+Z_{22}}\right) = R_{11} - \left\{\frac{(R_{12}^2 - X_{12}^2)(R+R_{22}) + 2R_{12}X_{12}(X+X_{22})}{(R+R_{22})^2 + (X+X_{22})^2}\right\},$$

$$\frac{|i_2|^2}{|i_1|^2} = \frac{(R_{12}^2 + X_{12}^2)}{(R+R_{22})^2 + (X+X_{22})^2}. \quad (I.64)$$

Substituting these values in (I.63), and combining terms, we finally have

$$\eta_c = \frac{R(R_{12}^2 + X_{12}^2)/R_{11}}{\left[R + R_{22} - \dfrac{R_{12}^2 - X_{12}^2}{2R_{11}}\right]^2 + \left[X + X_{22} - \dfrac{R_{12}X_{12}}{R_{11}}\right]^2 - \left[\dfrac{R_{12}^2 + X_{12}^2}{2R_{11}}\right]^2}. \quad (I.65)$$

While this is a rather formidable expression, it represents a function not greatly different in its properties from (I.59). We can show without trouble that the contours of constant circuit efficiency, in a Z plane, are circles, all orthogonal to the family of circles passing through the two points

$$R = \pm \left\{\left[R_{22} - \frac{R_{12}^2 - X_{12}^2}{2R_{11}}\right]^2 - \left[\frac{R_{12}^2 + X_{12}^2}{2R_{11}}\right]^2\right\}^{\frac{1}{2}}, \quad X = -X_{22} + \frac{R_{12}X_{12}}{R_{11}}. \quad (I.66)$$

The circuit efficiency has a maximum at the point corresponding to the positive sign, and decreases everywhere from there until it is zero on the axis of ordinates, or for a reactive load. We shall have occasion later to consider circuit efficiency, in connection with the power output of magnetrons, klystrons, and other microwave generators, and shall postpone more detailed discussion of these formulas until then, when we shall put them in simpler form.

17. Resumé of Network Theory

In this chapter we have given a discussion of some phases of network theory; we have omitted many important points, but have taken up those which we shall most particularly want to use in our later work. Before going on, we should emphasize that our results flow merely from the assumption (I.1) or (I.2) that the various voltages are linear functions of the corresponding currents. We have such relations for lumped constant networks, but we shall find as well that we have such relations for the oscillations of resonant cavities. We shall now proceed to derive the properties of wave guides and of resonant cavities from Maxwell's equations, and to show that we can define quantities analogous to voltage and current, which satisfy these same relations. We shall then be able to apply all the analysis of this chapter to the results of our theory. We shall not be merely using analogies with ordinary circuit theory; we shall be deriving results directly from the mathematical nature of our solutions. We turn in the next chapter to wave guides, and show that they form a perfect analogy to the continuous transmission lines of Section 9.

II. WAVE GUIDES

1. The Electromagnetic Field in a Wave Guide

By a wave guide we mean a cylindrical pipe, bounded by a conductor (of high conductivity), filled with a dielectric (of low loss). It may have arbitrary cross section; the two commonest cases in practice are the rectangular and circular guide. It may have more than one bounding surface, as the coaxial line, which consists of the annular space between two concentric circular cylindrical conductors. Its object is to transmit electromagnetic power. We shall first consider the guide without losses, either in the conducting walls or the dielectric, and shall show that in that case a disturbance can be propagated down the guide, closely analogous to the disturbance in a lossless continuous transmission line. Our problem is to find $\mathbf{E}$ and $\mathbf{H}$, solutions of Maxwell's equations, within the guide, satisfying suitable boundary conditions at the surface of the guide. Within a perfect conductor, no fields, electric or magnetic, can exist. Thus, since the normal component of $\mathbf{B}$ and the tangential component of $\mathbf{E}$ must be continuous at a surface, the boundary conditions are that $\mathbf{E}$ is normal, $\mathbf{H}$ tangential, to the surface. We shall now show that we can set up solutions of the problem of the form

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_0 \exp[j(\omega t - \beta z)], \\ \mathbf{H} &= \mathbf{H}_0 \exp[j(\omega t - \beta z)], \end{aligned} \tag{II.1}$$

where $\mathbf{E}_0$, $\mathbf{H}_0$ are vector functions of x and y, the coordinates in the plane of the cross section of the guide, which themselves satisfy the condition that $\mathbf{E}_0$ is normal, $\mathbf{H}_0$ tangential, to the surface. To prove this statement, we must show that the expressions (II.1) satisfy Maxwell's equations

$$\begin{aligned} \text{curl}\,\mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0, \quad \text{div}\,\mathbf{B} = 0, \\ \text{curl}\,\mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} = \mathbf{J}, \quad \text{div}\,\mathbf{D} = \rho \end{aligned} \tag{II.2}$$

where we assume that the guide is empty, so that within it we have

$$\mathbf{D} = \epsilon_0 \mathbf{E}, \quad \mathbf{B} = \mu_0 \mathbf{H}, \quad \mathbf{J} = 0, \quad \rho = 0. \tag{II.3}$$

We shall use rationalized m.k.s. units; that is, $\mathbf{E}$ is measured in volts per meter, $\mathbf{D}$ in coulombs per square meter, $\mathbf{H}$ in ampere-turns per meter, $\mathbf{B}$ in webers per square meter (one weber per square meter $= 10^4$ gauss), $\mathbf{J}$ in amperes per square meter, ρ in coulombs per cubic meter, distances in meters, and times in seconds. The quantities ϵ_0 and μ_0 are

$$\begin{aligned} \epsilon_0 &= 8.85 \times 10^{-12} \text{ farad per meter} \\ \mu_0 &= 4\pi \times 10^{-7} \text{ henry per meter} \end{aligned} \tag{II.4}$$

and satisfy the relations

$$\begin{aligned} (\mu_0/\epsilon_0)^{\frac{1}{2}} &= 376.6 \text{ ohms}, \\ (\epsilon_0\mu_0)^{-\frac{1}{2}} &= c = 3.00 \times 10^8 \text{ meter/sec.} \end{aligned} \tag{II.5}$$

From Maxwell's equations, in the usual way, we can derive the wave equations,

$$\nabla^2 \mathbf{E} - \frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} = 0, \quad \nabla^2 \mathbf{H} - \frac{1}{c^2}\frac{\partial^2 \mathbf{H}}{\partial t^2} = 0. \tag{II.6}$$

We shall first ask under what conditions the fields (II.1) will satisfy the wave equations. Substituting, we have at once

$$\frac{\partial^2 \mathbf{E}_0}{\partial x^2} + \frac{\partial^2 \mathbf{E}_0}{\partial y^2} + \left(\frac{\omega^2}{c^2} - \beta^2\right)\mathbf{E}_0 = 0 \tag{II.7}$$

with the same equation for $\mathbf{H}_0$. We shall rewrite this in the form

$$\frac{\partial^2 \mathbf{E}_0}{\partial x^2} + \frac{\partial^2 \mathbf{E}_0}{\partial y^2} + \left(\frac{2\pi}{\lambda_c}\right)^2 \mathbf{E}_0 = 0 \tag{II.8}$$

where we have

$$\frac{\omega^2}{c^2} = \left(\frac{2\pi}{\lambda_0}\right)^2 = \beta^2 + \left(\frac{2\pi}{\lambda_c}\right)^2 = \left(\frac{2\pi}{\lambda_g}\right)^2 + \left(\frac{2\pi}{\lambda_c}\right)^2$$

or

$$\frac{1}{\lambda_0^2} = \frac{1}{\lambda_g^2} + \frac{1}{\lambda_c^2}. \tag{II.9}$$

Here we have introduced a quantity λ_0, which is clearly the free space wave-length, or the wave-length which a disturbance with angular frequency ω would have if propagated in free space. We have also introduced λ_g, equal to $2\pi/\beta$. This quantity will be called the guide wave-length; from (II.1) we see that it is the wave-length with which the wave is propagated down the guide. The quantity λ_c will be called the cut-off wave-length. The reason for this is simple. If the free space wave-length is smaller than the cut-off wave-length, then $1/\lambda_0^2$ is greater than $1/\lambda_c^2$, so that (II.9) tells us that $1/\lambda_g^2$ is positive, and λ_g is real. On the other hand, if the free space wave-length becomes larger than the cut-off wave-length, $1/\lambda_g^2$ becomes negative, and the guide wave-length λ_g is imaginary. Inserting an imaginary wave-length into (II.1), we see that the solution represents an attenuated rather than a propagated wave. This corresponds to the case of the wave guide beyond cut-off mentioned in Section 9, Chapter I, and leads to no transfer of power down the line. Thus the guide acts like a high pass filter, only wave-lengths shorter (or with higher frequency) than the cut-off wave-length being propagated.

2. Transverse Electric and Transverse Magnetic Modes

The fields **E** and **H** must satisfy not only the wave equation, but Maxwell's equations as well. When we write these equations down, we find at once that two separate types of solutions are possible: solutions for which $E_z=0$, and solutions for which $H_z=0$. The first type has **E** transverse to the guide, and is called a transverse electric mode (abbreviated *TE*), and the second has **H** transverse, and is called transverse magnetic (*TM*). We shall let $\mathbf{E}_t$, $\mathbf{H}_t$, refer to the transverse components of **E** and **H**, and E_z and H_z be the components along the guide. Furthermore we let **k** be unit vector along the z direction, or along the guide. Then from Maxwell's equations we find directly that these components must satisfy the following relations:

$$TE:\quad \text{grad}\, H_z = 2\pi j\frac{\lambda_g}{\lambda_c^2}\mathbf{H}_t,$$
$$TM:\quad \text{grad}\, E_z = 2\pi j\frac{\lambda_g}{\lambda_c^2}\mathbf{E}_t. \qquad \text{(II.10)}$$

In these expressions, H_z and E_z represent the magnitudes of the corresponding vectors. Since they are functions of x and y, their gradients are in the xy plane, as is proper for $\mathbf{H}_t$ or $\mathbf{E}_t$. We find that there is a relationship between the transverse components of **E** and **H**, as follows:

$$\mathbf{H}_t=\frac{\mathbf{k}\times\mathbf{E}_t}{Z_0},\quad \text{where}\quad Z_0=\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}\frac{\lambda_g}{\lambda_0}\ \text{for } TE$$
$$=\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}\frac{\lambda_0}{\lambda_g}\ \text{for } TM. \qquad \text{(II.11)}$$

This means, since **k** and $\mathbf{E}_t$ are at right angles to each other, that $\mathbf{H}_t$ is at right angles to $\mathbf{E}_t$ in the xy plane, and is equal in magnitude to the magnitude of $\mathbf{E}_t$, divided by the quantity Z_0. Using (II.10) and (II.11), we can find all the components of **E** and **H** from H_z (in the *TE* case) or from E_z (in the *TM* case). These quantities, like the transverse components, satisfy the wave equation (II.8), and are scalar solutions of that differential equation. Furthermore, E_z is zero on the boundary of the guide (since E must have no tangential component on the surface), while H_z has a vanishing normal derivative on the boundary (since **H** must have no normal component on the surface, and $\mathbf{H}_t$ is proportional to the gradient of H_z). We may draw lines of $H_z=$ constant (in the *TE* case) or of $E_z=$ constant (in the *TM* case) in the xy plane. Then by (II.10) the orthogonal trajectories of these lines will be along the direction of $\mathbf{H}_t$ (in the *TE* case) or of $\mathbf{E}_t$ (in the *TM* case). Finally, since by (II.11) the direction of $\mathbf{H}_t$ is perpendicular to that of $\mathbf{E}_t$, the lines of constant H_z will be along the direction of $\mathbf{E}_t$ (in the *TE* case), and the lines of constant E_z will be along the direction of $\mathbf{H}_t$ (in the *TM* case). Proceeding in this way, we may draw lines of force, for the transverse components of **E** and **H**, in the xy plane, finding of course that the electric lines of force meet the surfaces at right angles, while the magnetic lines of force are tangential to the surface.

For every scalar solution of the two-dimensional wave equation satisfying the condition that it vanishes on the boundary, we get a *TM* wave, and for every solution whose normal derivative vanishes we get a *TE* wave, as we have seen above. There will be an infinite number of solutions of each type, each corresponding to a particular cut-off wave-length. These wave-lengths may be arranged in order of decreasing magnitude; they start with a largest cut-off wave-length, associated with the lowest mode of oscillation, and extend indefinitely toward shorter and shorter wave-lengths, so that we have an infinite number of modes of oscillation. For instance, in a rectangular guide of dimensions a, b, the cut-off wave-lengths are given by the formula

$$\lambda_c=[(m/2a)^2+(n/2b)^2]^{-\frac{1}{2}}, \qquad \text{(II.12)}$$

where m, n are integers. If the dimension a of the guide is greater than b, the longest cut-off wave-length is given by $m=1$, $n=0$, and is equal to $2a$. In general, it is more convenient to describe the modes of a guide by a single index n, which we shall take to be the number of the mode when arranged in order of descending wave-length. We shall denote the functions $\mathbf{E}_t$, E_z, $\mathbf{H}_t$, H_z, Z_0, λ_c, λ_g, for the nth mode by an additional subscript n.

In some cases, the first mode has an infinite cut-off wave-length; in this case we call it a principal mode. We should have such a case in (II.12) if m and n were both zero, but it turns out that this mode does not exist in this case, for the field described by these integers becomes equal to zero identically. Such a situation does not occur always, however. We find that a principal mode exists if the wall of the wave guide consists of at least two separated conductors, as for instance in the coaxial line. When a mode of infinite cut-off wave-length, or principal mode, exists, it has great practical importance, because it can be used to propagate any wave-length, no matter how long. The commonly used mode of the coaxial line is a principal mode, and the familiar parallel wire transmission line, ordinarily used for low frequencies, can also be considered as a wave guide with a principal mode. It is proved quite generally, on the other hand, that any wave guide whose wall consists of only one conductor has no principal mode. The physical reason for this is quite clear: we can put a very low frequency, or direct current, into a transmission line consisting of two or more conductors, and they will be insulated from each other, and suited to conduct the current. If there is only one conductor, however, as in an ordinary hollow pipe, there would clearly be a short circuit for a low frequency or direct current, and no propagation is possible until we get to a wave-length short enough so that something like real wave propagation occurs. For a principal mode, the cut-off wave-length is infinite, so that (II.10) tells us that grad H_z, or grad E_z, must be zero. That is, the longitudinal components of both E and H are zero, and such a wave is simultaneously transverse electric and transverse magnetic. It is sometimes called a transverse electromagnetic wave (*TEM*) for this reason. Furthermore for such a wave, as we see from (II.9), the guide wave-length becomes equal to the free space wave-length, so that the velocity of propagation becomes c. Finally, from (II.11), the quantity Z_0 for a principal wave becomes $(\mu_0/\epsilon_0)^{\frac{1}{2}}$, which can be shown to be the ratio of the magnitudes of E and H in a plane wave in free space. Thus a principal wave has many of the properties of a wave in free space.

At any given frequency, or free space wave-length λ_0, a given wave guide will in general support disturbances corresponding to all the modes. We now see, however, that some of these disturbances will be really propagated, but others will be attenuated. In fact, the only disturbances which are propagated will be those for the finite number of modes whose cut-off wave-lengths are greater than the free space wave-length; the remaining infinite number of modes with cut-off wave-lengths shorter than the free space wave-length will be attenuated. There will be a certain range of free space wave-lengths, between the longest and the next longest cut-off wave-lengths, in which only the lowest mode, often referred to as the dominant mode, will be propagated, and wave guides are generally used in this range, so as to avoid the difficulty of having many modes simultaneously present. Guides, in other words, are ordinarily used only over a rather limited range of wave-lengths. A coaxial line, on the contrary, is used in its principal mode, or for wave-lengths greater than the next cut-off wave-length beyond that of the principal mode. Thus the coaxial line is used as a low pass device, but the ordinary wave guide as a band pass device.

3. Standing Waves and Reflection Coefficients

In addition to the solution (II.1) of Maxwell's equations, corresponding to a wave traveling along the positive z direction, there is of course a wave traveling along the negative z direction, characterized by the opposite sign for β. Formally this brings about a change in the sign of λ_g, and hence, in (II.10) and (II.11), a change in the sign in the relation between $\mathbf{E}_t$ and $\mathbf{H}_t$, and in the relation between H_z and $\mathbf{H}_t$, or between E_z and $\mathbf{E}_t$. If we choose to keep the signs of (II.10) and (II.11), we must then change signs appropriately in writing the formulas for $\mathbf{E}$ and $\mathbf{H}$. Doing this, and superposing waves traveling in both directions, with appropriate amplitudes, we may write the field as

$$\mathbf{E}=\mathbf{E}_{tn}(A_n e^{j(\omega t-\beta_n z)}-B_n e^{j(\omega t+\beta_n z)})$$
$$+\mathbf{E}_{zn}(A_n e^{j(\omega t-\beta_n z)}+B_n e^{j(\omega t+\beta_n z)}),$$
$$\mathbf{H}=\mathbf{H}_{tn}(A_n e^{j(\omega t-\beta_n z)}+B_n e^{j(\omega t+\beta_n z)})$$
$$+\mathbf{H}_{zn}(A_n e^{j(\omega t-\beta_n z)}-B_n e^{j(\omega t\beta_n z)}) \quad \text{(II.13)}$$

where the relations between the various quantities are just as in (II.10) and (II.11). By comparison for instance with (I.11) we recognize the behavior of the transverse components of $\mathbf{E}$ and $\mathbf{H}$ as being analogous to the voltage and current in a transmission line. Just as in (I.31) we can introduce a reflection coefficient

$$r=-\frac{B_n}{A_n}e^{2j\beta_n z}. \quad \text{(II.14)}$$

(We must note that the subscript n here refers to the nth mode of the wave guide, a completely different meaning from that in Chapter I, where we were using it to refer to the nth set of terminals in a transmission line made of discrete four-terminal networks.) This reflection coefficient has the same properties as that of Chapter I: its magnitude represents the ratio of reflected amplitude to incident amplitude, and its phase, as we go along the line, rotates through an angle of 2π in a half-wave-length, assuming that β is real, or that we are dealing with real propagation. If on the other hand we have attenuation the transformation of going along the line corresponds to a shrinking of all points toward the origin in the reflection coefficient space. Mathematically, we can set up standing wave ratios in voltage, in power, and in decibels, as in (I.34), (I.35), and (I.36). We must ask how we measure these standing wave ratios, however. In the experimental measurement of standing waves in a guide, a small probe is inserted through a slot in the wall of the guide, the slot being located at a point in the cross section where no current must flow, so that the field inside the guide is not disturbed appreciably by the slot. The probe is connected to a coaxial line or other type of line terminated by a power measuring device. The amplitude of the wave set up in this coaxial line is proportional to the component of $\mathbf{E}$ along the direction of the probe, or transverse to the guide, and the power absorbed by the power measuring device is proportional to the square of the transverse component of $\mathbf{E}$. Thus $\mathbf{E}_t$ plays the same part in finding the power measured in standing wave measurements with a guide that the voltage does in the transmission line of Chapter I, as discussed in Section 10.

4. Impedance and Power Flow

In Chapter I, for instance in Eq. (I.37), we have found the relation between impedance and reflection coefficient:

$$Z/Z_0=(1+r)/(1-r). \quad \text{(II.15)}$$

This defines the ratio of impedance to characteristic impedance uniquely in terms of reflection coefficient. Using this definition, we can introduce a ratio of impedance to characteristic impedance uniquely for a wave guide. We shall refer to this ratio as the reduced impedance. The properties of it are of course just as in Chapter I, since it is defined in the same way. There is, however, no unique way of defining the characteristic impedance for a wave guide, and hence no unique way of defining the impedance itself. From (II.13), we may take the ratio of the magnitude of $\mathbf{E}_t$ to the magnitude of $\mathbf{H}_t$ (where we mean magnitude in the sense of three-dimensional vectors in space, not in the sense of complex vectors representing the sinusoidal time variation). Using (II.11), we have

$$|E_t|/|H_t|=Z_0(1+r)/(1-r), \quad \text{(II.16)}$$

where Z_0 now is as defined in (II.11). We see, in other words, that if we choose to interpret that quantity as a characteristic impedance, the ratio of magnitudes of tangential components of $\mathbf{E}$ and $\mathbf{H}$ will play the part of an impedance. On the other hand, if we choose to use any multiple of the quantity Z_0 of (II.11) as the characteristic impedance, and the corresponding multiple of the ratio of $\mathbf{E}_t$ to $\mathbf{H}_t$ as impedance, we shall equally well have agreement with (II.15). We shall use the quantity (II.11) as a characteristic impedance; for as we shall show shortly, there is no other definition of characteristic impedance of a wave guide which is more universally sensible.

For comparison with the results of Chapter I, we should consider not merely the impedance, but the power flow as well. This is of course uniquely determined, since it can be directly measured, by terminating the guide by a power measuring device, as a bolometer or thermistor or water load, whose temperature rise indicates the total power absorbed by it, and by assuming that all power flowing down the guide is absorbed in the power measurer. Mathematically, we can find the power flow by integrating the normal component of Poynting's vector over a cross section of the guide. Poynting's vector is $\mathbf{E}\times\mathbf{H}$, and its normal component, or z component, is $\mathbf{k}\cdot(\mathbf{E}\times\mathbf{H})$. The time average is easily seen to be computed as the time average power was in Chapter I, Section 14: it is $\frac{1}{2}Re\mathbf{k}\cdot(\mathbf{E}\times\bar{\mathbf{H}})$. In Poynting's vector we encounter the vector quantity $\mathbf{k}\cdot(\mathbf{E}_t\times\bar{\mathbf{H}}_t)$: the other quantities $\mathbf{k}\cdot(\mathbf{E}_z\times\bar{\mathbf{H}}_t)$, $\mathbf{k}\cdot(\mathbf{E}_t\times\bar{\mathbf{H}}_z)$, and $\mathbf{k}\cdot(\mathbf{E}_z\times\bar{\mathbf{H}}_z)$ are all automatically zero. We notice that as a result of (II.11) we have

$$\mathbf{k}\cdot(\mathbf{E}_t\times\bar{\mathbf{H}}_t)=\frac{|E_t|^2}{Z_0}=Z_0|H_t|^2. \quad \text{(II.17)}$$

Using these relations, we may then write the time average Poynting's vector, S, in the form appropriate for the case of real propagation,

$$S=\frac{1}{2}\frac{|E_{tn}|^2}{Z_{0n}}(A_n\bar{A}_n-B_n\bar{B}_n)$$
$$=\tfrac{1}{2}Z_{0n}|H_t|^2(A_n\bar{A}_n-B_n\bar{B}_n). \quad \text{(II.18)}$$

To get the total power flowing through the guide, we must integrate this quantity over the area of the guide, so that we have

$$P=\frac{1}{2}\frac{\int|E_{tn}|^2da}{Z_{0n}}(A_n\bar{A}_n-B_nB_n),$$
$$=\frac{1}{2}Z_{0n}\int|H_{tn}|^2da(A_n\bar{A}_n-B_n\bar{B}_n). \quad \text{(II.19)}$$

We have so far left the values of the integrals of $|E_{tn}|^2$ or $|H_{tn}|^2$ over the guide arbitrary; obviously in (II.13) we can multiply the quantities E_{tn} and H_{tn} by an arbitrary constant, and divide A_n and B_n by the same constant, without changing E and H, which alone have physical significance. The value which we choose for these constants is purely a matter of convenience. Since we have already decided to make Z_0 analogous to a characteristic impedance, comparison with (I.55) suggests that we make the A_n's and B_n's analogous to the A's and B's introduced in Chapter I, which by (I.53) are current amplitudes. To accomplish this, we may assume

$$\int|H_{tn}|^2da=1, \quad \text{(II.20)}$$

so that

$$P=\tfrac{1}{2}Z_{0n}(A_n\bar{A}_n-B_nB_n). \quad \text{(II.21)}$$

It is now clear that our values of impedance and power flow in a wave guide are analogous to the corresponding quantities for a transmission line, if we determine the magnitude of H_{tn} by (II.20) (called a normalization condition), and if we postulate a voltage V_n and current i_n given by

$$V_n=Z_{0n}A_ne^{j(\omega t-\beta_n z)}-B_ne^{j(\omega t+\beta_n z)},$$
$$i_n=A_ne^{j(\omega t-\beta_n z)}+B_ne^{j(\omega t+\beta_n z)} \quad \text{(II.22)}$$

so that V_n is proportional to the transverse $\mathbf{E}$, i_n to the transverse $\mathbf{H}$, in the nth mode. Since these equations are entirely analogous to those of Chapter I, we have mathematically justified the results of that chapter, as applied to the wave guides. By entirely similar methods, we can justify formulas like (I.57) with corresponding interpretation of voltage and current, for the case where the guide is beyond cut-off, so that there is only attenuation, not propagation, and where the characteristic impedance is pure imaginary.

While we have suggested a particular way of setting up a voltage and current, this is by no means the only possible way. In fact, it is obvious that we can assume that the voltage is any constant times the transverse $\mathbf{E}$, and the current any constant times the transverse $\mathbf{H}$. This gives two arbitrary constants in the interpretation of the behavior of the guide as a transmission line. On the other hand, if the power is to be determined by the relation $P=\frac{1}{2}Re\,V\bar{i}$, this imposes a relation between the definitions of voltage and currents, so that only one arbitrary constant is left. We may still use this arbitrary constant to make the characteristic impedance, or the voltage, or the current, anything we please, but if one of these quantities is determined, the others are also. In a few cases there is an obvious way to define voltage and current. For instance, in a coaxial line, or other line possessing a principal mode, the voltage is uniquely defined: for in that case (II.7) shows that the transverse **E** obeys Laplace's equation, so that its integral from one conductor to the other, being independent of path, forms a unique voltage. Similarly there is a unique current, the actual current flowing in either conductor. The voltage and current so defined do not agree with our value (II.22). Again, in the lowest mode of a rectangular wave guide, reasonable definitions of voltage and current can be given. These definitions again do not agree with (II.22), but neither do they have a simple relation to those used for coaxial lines. In fact, it seems to be impossible to set up any general definition of voltage and current which reduces in a reasonable way to the natural values met in these simple cases. For that reason we adopt our definitions (II.22), which are the simplest ones. It is a fortunate fact that this ambiguity in the definition of current, voltage, and impedance really does not affect us at all; for the quantities which actually are important are the ratio of impedance to characteristic impedance, or reduced impedance, and the power flow, which are uniquely determined, quite apart from this ambiguity.

5. Expansion of the Field in Normal Modes

So far, we have assumed that the field consists of a single normal mode only, but of course on account of the linear nature of Maxwell's equations the general solution of our problem is a superposition of all normal modes, each with its appropriate amplitude and phase; that is, **E** and **H** are given by summations over n of the quantities given in (II.13), rather than just the nth term. In the present section we shall take up those properties of the field that depend on the fact that ordinarily all modes are simultaneously excited. The discussion is made possible by the proof of several theorems related to the orthogonality of the normal modes. We shall first prove these theorems. They are as follows:

(I) The integral over the cross section of the guide of the scalar product $\mathbf{E}_{tn}\cdot\mathbf{E}_{tm}$, or $\mathbf{H}_{tn}\cdot\mathbf{H}_{tm}$, where n and m are different, is zero.

(II) The integral over the cross section of the guide of the product $E_{zn}E_{zm}$, or $H_{zn}H_{zm}$, where n and m are different, is zero.

(III) The integral over the cross section of the guide of the quantity $\mathbf{k}\cdot(\mathbf{E}_{tn}\times\mathbf{H}_{tm})$, where n and m are different, is zero.

The proofs follow easily from two-dimensional forms of Green's theorem. We first use Green's theorem in the form

$$\int(\phi\nabla^2\psi+\text{grad}\,\phi\cdot\text{grad}\,\psi)da=\int\phi\frac{\partial\psi}{\partial n}ds,\qquad\text{(II.23)}$$

where the integral on the left is over an area (in this case the area of the cross section), the integral on the right is over the perimeter, and $\partial\psi/\partial n$ is the normal derivative in the direction of the outer normal to the area, and where ϕ, ψ, are two scalar functions of position. We let ϕ be E_{zn}, ψ be E_{zm}, and remember that on account of (II.8) we have

$$\nabla^2E_{zn}+(2\pi/\lambda_{cn})^2E_{zn}=0.\qquad\text{(II.24)}$$

Thus we have, using (II.10),

$$-\left(\frac{2\pi}{\lambda_{cn}}\right)^2\int E_{zn}E_{zm}da-4\pi^2\frac{\lambda_{gn}\lambda_{gm}}{\lambda_{cn}{}^2\lambda_{cm}{}^2}\int\mathbf{E}_{tn}\cdot\mathbf{E}_{tm}da$$
$$=\int E_{zn}(\mathbf{n}\cdot\text{grad}\,E_{zm})ds.\qquad\text{(II.25)}$$

Since $E_{zn}=0$ on the perimeter, the line integral on the right is zero, and we thus see that if $\int E_{zn}E_{zm}da=0, \int\mathbf{E}_{tn}\cdot\mathbf{E}_{tm}da=0$ as well. That is, we show that theorems (I) and (II) are equivalent, as far as **E** is concerned. To show the same thing for the H's, we proceed in the same way but now the line integral in (II.25) vanishes because grad H_{zm} is parallel to the surface, since $\mathbf{H}_{tm}$ has no component normal to the surface. Next we use Green's theorem in the form

$$\int(\phi\nabla^2\psi-\psi\nabla^2\phi)da=\int\left(\phi\frac{\partial\psi}{\partial n}-\psi\frac{\partial\phi}{\partial n}\right)ds,\qquad\text{(II.26)}$$

and let $\phi=E_{zn}$, $\psi=E_{zm}$. The right-hand side vanishes as before, and the left-hand side gives

$$\left(\frac{1}{\lambda_{cn}^2}-\frac{1}{\lambda_{cm}^2}\right)\int E_{zn}E_{zm}da=0, \quad (II.27)$$

so that, if n and m are different, and the first factor is not zero, we have $\int E_{zn}E_{zm}da=0$, proving theorem (II), and hence theorem (I). The corresponding proof for the H's follows similarly. To prove (III), we need only note that, by (II.11), $\mathbf{H}_{tm}=(\mathbf{k}\times\mathbf{E}_{tm})/Z_{0m}$, from which we show at once that

$$\mathbf{k}\cdot(\mathbf{E}_{tn}\times\mathbf{H}_{tm})=(\mathbf{E}_{tn}\cdot\mathbf{E}_{tm})/Z_{0m}, \quad (II.28)$$

so that (III) reduces to (I).

We shall first use these theorems to prove that when we consider the flow of energy, or the total energy in the guide, we may handle the various modes separately, the total energy flow or energy being the sum of the corresponding terms for the various modes, without cross terms. For the energy flow, we must compute the integral over the cross section of the quantity $\frac{1}{2}Re\mathbf{k}\cdot(\mathbf{E}\times\bar{\mathbf{H}})$, where $\mathbf{E}$ and $\mathbf{H}$ are given by summations over n of the terms as in (II.13). This will involve terms $\mathbf{k}\cdot(\mathbf{E}_{tn}\times\mathbf{H}_{tm})$, whose integrals over the cross section are zero, by (III), if n and m are different. The only non-vanishing terms are then those for $n=m$, and these are simply the terms computed for the various modes separately, as in Section 4. Thus the total flow of power is the sum of the flows of the various modes. Similarly the total energy is the integral over the volume of $\frac{1}{4}Re(\epsilon_0\mathbf{E}\cdot\bar{\mathbf{E}}+\mu_0\mathbf{H}\cdot\bar{\mathbf{H}})$. The quantity $\mathbf{E}\cdot\bar{\mathbf{E}}$ or $\mathbf{H}\cdot\bar{\mathbf{H}}$ is again the product of two sums from (II.13), and we again write it as a double sum. The terms involved are of the form $\mathbf{E}_{tn}\cdot\mathbf{E}_{tm}$, $\mathbf{E}_{zn}\cdot\mathbf{E}_{zm}$, $\mathbf{H}_{tn}\cdot\mathbf{H}_{tm}$, or $\mathbf{H}_{zn}\cdot\mathbf{H}_{zm}$, all of which integrate to zero over the cross section of the guide, as we see from theorems (I) and (II), if n and m are different. The only remaining terms again are those for $n=m$, which are the terms relating to the individual modes, so that the total energy is the sum of the energies of the modes, without cross terms. The superposition of modes, in other words, brings about no complications when we consider energy and its flow.

The more interesting application of our theorems I to III comes in setting up the expansion of the field in the guide, subject to certain boundary conditions. We may well ask, how much information about the field in a guide is necessary to determine it uniquely? The answer is simple: we must know the tangential components of $\mathbf{E}$ and $\mathbf{H}$ as functions of x and y across a single cross section. We shall prove this by setting up the expansion coefficients uniquely in terms of that information. For the sake of simplicity, we shall choose the plane on which we know the tangential values of $\mathbf{E}$ and $\mathbf{H}$ as the plane $z=0$. In that plane, then, using (II.13), the tangential components of $\mathbf{E}$ and $\mathbf{H}$ are

$$\begin{aligned}\mathbf{E}_t e^{j\omega t}&=\textstyle\sum_n \mathbf{E}_{tn}(A_n+B_n)e^{j\omega t},\\ \mathbf{H}_t e^{j\omega t}&=\textstyle\sum_n \mathbf{H}_{tn}(A_n-B_n)e^{j\omega t}.\end{aligned} \quad (II.29)$$

In this expression, $\mathbf{E}_t$ and $\mathbf{H}_t$ are assumed to be known functions of x and y, the factors $e^{j\omega t}$ expressing their variation with time. If the field is not sinusoidal with time, we first make a Fourier integral analysis in time, and apply our discussion to a single sinusoidal component. We now take the first equation of (II.29), take its scalar product with one of the $\mathbf{E}_{tm}$'s, and integrate ever the cross section of the guide. Using our theorem (I), all terms except the mth will drop out. We may then use the normalization condition

$$\int|E_{tn}|^2da=Z_{0n}^2, \quad (II.30)$$

which follows from (II.17) and (II.20), and find

$$\frac{\int\mathbf{E}_t\cdot\mathbf{E}_{tn}da}{Z_{0n}^2}=A_n+B_n. \quad (II.31)$$

Similarly we multiply the second equation of (II.29) by one of the $\mathbf{H}_{tm}$'s and integrate. Using the normalization condition (II.20), we have

$$\int\mathbf{H}_t\cdot\mathbf{H}_{tn}da=A_n-B_n. \quad (II.32)$$

From (II.31) and (II.32) we can find A_n and B_n in terms of integrals of the known functions $\mathbf{E}_t$ and $\mathbf{H}_t$. Thus we can set up the summation of terms (II.13), and hence the complete field within the guide, showing that the tangential component of field over a single cross section determines the complete field.

6. Losses in the Wave Guide

In our treatment so far, we have assumed that the walls of the wave guide were perfect conductors, so that **E** had to be normal to the walls. In that case there is no power flow into the walls, and the guide is a lossless transmission line. If the walls have a finite conductivity, however, power will be dissipated in them, and the guide will show attenuation. From the theory of the skin effect, which we shall not go into,* we find that a disturbance of angular frequency ω, in a good conductor of conductivity σ, with the same magnetic permeability μ_0 as free space, penetrates only a short distance into the conductor; the amplitude of both electric and magnetic fields falls to $1/e$ of the value on the surface in a distance δ equal to

$$\delta=(2/\sigma\mu_0\omega)^{\frac{1}{2}}. \qquad \text{(II.33)}$$

We notice that as the conductivity becomes infinite, the distance δ, sometimes called the skin depth, becomes zero, so that we approach the case we have treated earlier. We also find that there is a small tangential component of **E** at the surface, proportional to the tangential component of **H**, which measures the surface current, and at right angles to it, or in the same direction as the surface current. If **n** is the outer normal to the guide, we find that the tangential component of **E** at the surface is given in terms of the tangential **H** by the equation

$$\mathbf{E}=(\mathbf{H}\times\mathbf{n})(\mu_0\omega/2\sigma)^{\frac{1}{2}}(1+j), \qquad \text{(II.34)}$$

so that as the conductivity becomes infinite, the tangential **E** goes to zero. As a result of (II.34), there is a flow of energy into the surface, which we can find by computing the normal component of Poynting's vector. The time average is

$$\tfrac{1}{2}Re\,\mathbf{n}\cdot(\mathbf{E}\times\bar{\mathbf{H}})=\tfrac{1}{2}(\mu_0\omega/2\sigma)^{\frac{1}{2}}|H|^2. \qquad \text{(II.35)}$$

For a given tangential **H**, this flow of energy goes to zero as the conductivity becomes infinite. We can now use this discussion to give an approximate treatment of the effect of surface losses on the behavior of a wave guide. Clearly, with the boundary conditions (II.34) at the surface of the guide, the solutions for E and H which we have found in the preceding sections of this chapter are not correct. However, when we insert numerical values for ordinary metallic conductors, we find that the tangential E required at the surface is so small that the correct field is a very small perturbation of the field we have calculated earlier. Thus we are justified to a first approximation in assuming that the tangential H which we have found is that actually present, and that it can be used in computing the energy loss (II.35).

We may now, by simple calculation of power, find an approximate value of the attenuation constant α_n which is present in the nth mode, on account of the losses in the walls. In an attenuated wave, traveling along the positive z direction, the magnetic field **H**, by analogy with (II.13), will be

$$\mathbf{H}=(\mathbf{H}_{tn}+\mathbf{H}_{zn})A_n e^{j(\omega t-\beta_n z)}e^{-\alpha_n z}. \qquad \text{(II.36)}$$

By analogy with (II.19), the power flow will be

$$P=\tfrac{1}{2}Z_{0n}A_n\bar{A}_n e^{-2\alpha_n z}, \qquad \text{(II.37)}$$

in which we have also used (II.20). The loss of power into the walls, in unit length of the guide, will be $-(dP/dz)$. By (II.35), this will be

$$-\frac{dP}{dz}=\frac{1}{2}\left(\frac{\mu_0\omega}{2\sigma}\right)^{\frac{1}{2}}\int[|H_{tn}|^2+|H_{zn}|^2]ds \times A_n\bar{A}_n e^{-2\alpha_n z}, \qquad \text{(II.38)}$$

in which the line integral is to be taken around the perimeter of the cross section of the guide. At the same time, using (II.37), we have

$$-dP/dz=2\alpha_n P. \qquad \text{(II.39)}$$

Using (II.37), (II.38), and (II.39), we then find

$$\alpha_n=\tfrac{1}{2}(1/Z_{0n})(\mu_0\omega/2\sigma)^{\frac{1}{2}} \times\int[|H_{tn}|^2+|H_{zn}|^2]ds. \qquad \text{(II.40)}$$

This gives us a formula for α_n in terms of an integral of the square of the tangential component of H_n around the surface of the guide. This, it should be repeated, is correct only if we assume that the tangential **H** in the presence of

* See Slater, *Microwave Transmission*, Section 12, for a detailed discussion.

absorption is almost equal to that in the lossless case, for which it is assumed that it is calculated. We shall not give detailed examples of the application of this formula, but such examples are worked out in the various texts on microwaves. The one fact is obvious, that a guide in which the tangential H (and therefore the surface current) rises to high values at some points in the metallic surface will have high losses.

In this calculation of α_n from the power flow, we have assumed that one mode only was excited, and that there was only a direct wave, not a reflected wave. We may not assume here, however, that if we have a number of modes coexisting, the losses are simply a sum of the losses in the various modes. The integral of the square of the tangential **H** over the surface of the guide, which we find in (II.40), has no orthogonality property, and there is no way of disentangling the effects of the various modes. The losses, being a quadratic rather than a linear function of the amplitudes, have no principle of superposibility, and the presence of one mode can affect the losses experienced by another mode propagated through the same guide with the same frequency. The detailed study of this situation would be necessary to find the behavior in any particular case.

III. RESONANT CAVITIES

1. Orthogonal Functions for a Hollow Cavity

Just as a wave guide forms the microwave analogy for the transmission line of ordinary circuit theory, so a hollow cavity forms the analogy for a circuit element. A cavity can be provided with one or more output leads, in the form of wave guides of some type: rectangular guides, coaxial lines, etc. If it has only one lead, it serves as an impedance terminating that lead; if it has two, it serves as a transformer or transducer, allowing powder to flow into one lead, out the other. In the present chapter we take up the general theory of resonant cavities and the electromagnetic fields within them. We consider the case of an arbitrary number of output leads, and find the relation between the electromagnetic fields in these various leads, each of the form taken up in Chapter I. We shall be able to use these results later in discussing transformers and transducers, oscillators, and in fact all types of microwave problems.

Our first problem will be to solve Maxwell's equations in a hollow cavity, subject to certain boundary conditions around the surface. As with the problem of the wave guide which we have taken up in Chapter II, this solution will be in terms of a summation over certain normal modes, which possess orthogonality properties. The details of the process are quite different from the case of Chapter II, however, and we shall start from the beginning with our discussion. We shall start by postulating the properties of the normal functions, shall expand the electric and magnetic fields in terms of them, and shall then find what conditions must be satisfied to solve Maxwell's equations. We wish to solve Maxwell's equations within a volume bounded by a certain surface (nothing in our treatment will prevent this surface consisting of several parts, as the inner and outer surface of a hollow spherical shell). We shall find that we can set up orthogonal functions for two types of boundary conditions: the first, which we may call short circuited boundary conditions, requires that the tangential component of **E**, and the normal component of **H**, be zero on the surface, while the second, which we call open circuited boundary conditions, requires that the normal component of **E**, and the tangential component of **H**, be zero on the surface. The reason for the names is simple: a perfect conductor has zero tangential component of **E**, and forms the analog of a short circuit; while a perfect insulator carries no surface current, and hence, if **H** is zero within it, it demands a zero tangential component of **H**, and forms the analog of an open circuit. We shall discuss these points more in detail later. We shall find that we can use mixed boundary conditions, and that on occasion we shall want to: over part of the surface (which we denote by S) we shall have short circuited boundary conditions, while over the rest of the surface (which we denote by S') we shall have open circuited boundary conditions. Our object is now to set up orthogonal functions within the volume V bounded by S and S', suitable for expanding our fields within the volume.

Our first step is to notice that by general principles of vector analysis, any vector field can be broken up into two fields, one of which is solenoidal, or has zero divergence, and the other of which is irrotational, or has zero curl. We shall consequently set up two sets of orthogonal functions, one set solenoidal, the other irrotational, using the solenoidal functions to expand the solenoidal part of any vector field, and the irrotational functions to expand the irrotational part of a field. As a matter of fact, we go further: we set up two independent sets of solenoidal functions, one adapted for expanding the solenoidal part of **E**, the other adapted for expanding **H** (which is itself solenoidal, if we assume, as we shall, that the magnetic permeability μ_0 is a constant). We use only one type of irrotational function, used for expanding the irrotational part of **E**, and do not need another type only because **H** has no irrotational part. We shall denote the solenoidal functions used in expanding **E** by the symbol $\mathbf{E}_a$, and the solenoidal functions used in expanding **H** by the symbol $\mathbf{H}_a$. Similarly we shall denote the irrotational functions used in expanding **E** by the symbol $\mathbf{F}_a$. We shall now set up the equations used in defining these functions.

The functions $\mathbf{E}_a$ and $\mathbf{H}_a$, having no divergence, must be the curls of certain other vector functions, and we assume that they satisfy the equations

$$k_a\mathbf{E}_a = \text{curl } \mathbf{H}_a, \quad k_a\mathbf{H}_a = \text{curl } \mathbf{E}_a, \quad \text{(III.1)}$$

where k_a is a constant, which will later prove to be the propagation constant (2π divided by the wave-length) associated with the ath mode. We assume that $\mathbf{E}_a$ and $\mathbf{H}_a$ furthermore satisfy the following boundary conditions:

$$\mathbf{n}\times\mathbf{E}_a = 0 \text{ on } S, \quad \mathbf{n}\times\mathbf{H}_a = 0 \text{ on } S' \quad \text{(III.2)}$$

where **n** is the outer normal to the surface; that is, $\mathbf{E}_a$ has no tangential component over S, and $\mathbf{H}_a$ has no tangential component over S'. From (III.1) and (III.2), and Stokes' theorem, we can then prove that

$$\mathbf{n}\cdot\mathbf{H}_a = 0 \text{ on } S, \quad \mathbf{n}\cdot\mathbf{E}_a = 0 \text{ on } S'. \quad \text{(III.3)}$$

To prove the first of these, we take a small closed curve lying in the plane of the surface S, and integrate the line integral of the tangential component of $\mathbf{E}_a$ around this contour, which is zero since the tangential component of $\mathbf{E}_a$ is zero on S according to (III.2). By Stokes' theorem this line integral equals the surface integral of the normal component of curl $\mathbf{E}_a$, or of $k_a\mathbf{H}_a$, which is then zero, which is impossible, since we are dealing with an arbitrary contour, unless the normal component of $\mathbf{H}_a$, $\mathbf{n}\cdot\mathbf{H}_a$, is zero on S. The second statement of (III.3) is proved in a similar way.

We may easily set up separate differential equations for $\mathbf{E}_a$ and $\mathbf{H}_a$, instead of having them defined in terms of each other as in (III.1). We see at once that these equations are

$$\text{curl curl } \mathbf{E}_a = k_a^2\mathbf{E}_a, \quad \text{curl curl } \mathbf{H}_a = k_a^2\mathbf{H}_a. \quad \text{(III.4)}$$

Using the vector identity that curl curl $\mathbf{A}$ $=$ grad div $\mathbf{A} - \nabla^2\mathbf{A}$, and that div $\mathbf{E}_a = 0$, div $\mathbf{H}_a$ $= 0$, these become the familiar wave equation

$$\nabla^2\mathbf{E}_a + k_a^2\mathbf{E}_a = 0, \quad \nabla^2\mathbf{H}_a + k_a^2\mathbf{H}_a = 0. \quad \text{(III.5)}$$

These equations may be assumed to have an infinite set of solutions, corresponding to different values of k_a, subject to the boundary conditions (III.2). It is clear that corresponding to each k_a we have both a function $\mathbf{E}_a$ and an $\mathbf{H}_a$; both sets of functions correspond to the same set of characteristic numbers.

We shall now prove that the functions $\mathbf{E}_a$ and $\mathbf{H}_a$ have orthogonality properties of the form

$$\int_V \mathbf{E}_a\cdot\mathbf{E}_b dv = 0 \quad \text{if} \quad a\neq b,$$
$$\int_V \mathbf{H}_a\cdot\mathbf{H}_b dv = 0 \quad \text{if} \quad a\neq b. \quad \text{(III.6)}$$

To prove the first, we use the vector identity

$$\begin{aligned}\text{div } (\mathbf{E}_b\times\text{curl } \mathbf{E}_a) - \text{div } (\mathbf{E}_a\times\text{curl } \mathbf{E}_b) &= \text{curl } \mathbf{E}_a\cdot\text{curl } \mathbf{E}_b - \mathbf{E}_b\cdot\text{curl curl } \mathbf{E}_a \\ &\quad - \text{curl } \mathbf{E}_b\cdot\text{curl } \mathbf{E}_a + \mathbf{E}_a\cdot\text{curl curl } \mathbf{E}_b. \quad \text{(III.7)}\end{aligned}$$

Cancelling terms, and using (III.4), we can rewrite (III.7) in the form

$$\text{div } (\mathbf{E}_b\times\text{curl } \mathbf{E}_a) - \text{div } (\mathbf{E}_a\times\text{curl } \mathbf{E}_b) = (k_b^2 - k_a^2)\mathbf{E}_a\cdot\mathbf{E}_b. \quad \text{(III.8)}$$

Integrating over the volume V, and using Green's theorem to convert the left side into a surface integral, we have

$$\int_{S,S'} \mathbf{n}\cdot(k_a\mathbf{E}_b\times\mathbf{H}_a-k_b\mathbf{E}_a\times\mathbf{H}_b)da = (k_b^2-k_a^2)\int_V \mathbf{E}_a\cdot\mathbf{E}_b dv. \quad \text{(III.9)}$$

We may now easily show that the surface integral on the left vanishes on account of the boundary conditions. Over S, we may rewrite the integrand in the form $k_a\mathbf{H}_a\cdot(\mathbf{n}\times\mathbf{E}_b)-k_b\mathbf{H}_b\cdot(\mathbf{n}\times\mathbf{E}_a)$, which is zero on account of (III.2), while over S' we may rewrite it in the form $k_b\mathbf{E}_a\cdot(\mathbf{n}\times\mathbf{H}_b)-k_a\mathbf{E}_b\cdot(\mathbf{n}\times\mathbf{H}_a)$, which is likewise zero. Thus, if $k_b^2-k_a^2$ is different from zero, which will be the case if $a\neq b$ except in case of degeneracy, we have

$$\int_V \mathbf{E}_a\cdot\mathbf{E}_b dv=0 \quad \text{if} \quad a\neq b, \quad \text{(III.10)}$$

which we wished to prove. In the case of degeneracy we can prove, as in quantum mechanics, that we can always introduce normal functions $\mathbf{E}_a$ and $\mathbf{E}_b$ in such a way as to secure the orthogonality we desire, though it is no longer necessary that the functions have that property. The proof of the orthogonality of the $\mathbf{H}_a$'s, stated in (III.6), follows in an entirely analogous manner.

In addition to the orthogonality, we shall assume that the $\mathbf{E}_a$'s and $\mathbf{H}_a$'s are normalized in such a way that $\int E_a^2dv$ and $\int H_a^2dv$ equal unity, so that the normalization and orthogonality conditions can be written in the form

$$\int_V \mathbf{E}_a\cdot\mathbf{E}_b dv=\delta_{ab}, \quad \int_V \mathbf{H}_a\cdot\mathbf{H}_b dv=\delta_{ab}, \quad \text{(III.11)}$$

where as usual δ_{ab} is unity if $a=b$, zero if $a\neq b$. Since $\mathbf{E}_a$ and $\mathbf{H}_a$ are related by (III.1), we may not simultaneously assume that $\mathbf{E}_a$ and $\mathbf{H}_a$ are normalized in this way, without proof that the two conditions are consistent with each other. To prove this consistency, we have

$$\begin{aligned}\text{div}\,(\mathbf{E}_a\times\text{curl}\,\mathbf{E}_a) &= (\text{curl}\,\mathbf{E}_a)^2-\mathbf{E}_a\cdot\text{curl curl}\,\mathbf{E}_a \\ &= k_a^2(H_a^2-E_a^2). \end{aligned} \quad \text{(III.12)}$$

Integrating over V, the left side again transforms into a surface integral which vanishes, showing at once that $\int_V H_a^2dv=\int_V E_a^2dv$, so that each can be set equal to unity.

We shall next set up our functions $\mathbf{F}_a$, which have zero curl. On account of that property, they can be set equal to the gradient of a scalar function. We write

$$k_a\mathbf{F}_a=\text{grad}\,\psi_a. \quad \text{(III.13)}$$

We assume that the scalar ψ_a satisfies the wave equation

$$\nabla^2\psi_a+k_a^2\psi_a=0, \quad \text{(III.14)}$$

from which we immediately see that $\mathbf{F}_a$ also satisfies the wave equation. As boundary conditions we assume

$$\begin{aligned}\psi_a&=0 \text{ on } S \text{ and } S',\\ \mathbf{n}\times\mathbf{F}_a&=0 \text{ on } S \text{ and } S'.\end{aligned} \quad \text{(III.15)}$$

These conditions, in which the second obviously follows from the first, are not the most general boundary conditions which can be applied to these functions, but they will be sufficient for our purposes.

We can now prove that the functions $\mathbf{F}_a$ and ψ_a have orthogonality properties of the form

$$\begin{aligned}\int_V \mathbf{F}_a\cdot\mathbf{F}_b dv&=0 \quad \text{if} \quad a\neq b,\\ \int_V \psi_a\psi_b dv&=0 \quad \text{if} \quad a\neq b.\end{aligned} \quad \text{(III.16)}$$

To prove this,we have

$$\begin{aligned}\text{div}\,(\psi_a\,\text{grad}\,\psi_b) &= \psi_a\nabla^2\psi_b+\text{grad}\,\psi_a\cdot\text{grad}\,\psi_b\\ &= -k_b^2\psi_a\psi_b+\text{grad}\,\psi_a\cdot\text{grad}\,\psi_b.\end{aligned} \quad \text{(III.17)}$$

Interchanging the order of a and b, we set up another equation like (III.17), and subtract. We then integrate over V, obtaining

$$\int_{S,S'}\left(\psi_a\frac{\partial\psi_b}{\partial n}-\psi_b\frac{\partial\psi_a}{\partial n}\right)da = (k_a^2-k_b^2)\int_V \psi_a\psi_b dv. \quad \text{(III.18)}$$

The surface integral vanishes on account of the condition (III.15), so that, if $k_a\neq k_b$, the volume integral must vanish, leading to the orthogonality condition (III.16) for the ψ's. If we integrate (III.17) itself over V, the surface integral again vanishes, so that, if $\int_V \psi_a\psi_b dv=0$, we must also have $\int_V \text{grad}\, \psi_a\cdot\text{grad}\, \psi_b dv=0$, so that we prove the orthogonality of the $\mathbf{F}$'s. Finally we assume normalization of the form $\int_V F_a^2 dv=\int_V \psi_a^2 dv=1$, so that we can write the normalization and orthogonality in the form

$$\int_V \mathbf{F}_a\cdot\mathbf{F}_b dv=\int_V \psi_a\psi_b dv=\delta_{ab}. \quad \text{(III.19)}$$

To prove the consistency of these two conditions, we take (III.17), set $a=b$, and integrate over V. Using (III.13), we find at once that $\int_V \psi_a^2 dv =\int_V F_a^2 dv$, so that we are justified in setting them each equal to unity.

As a final step in setting up our orthogonal functions, we prove that one of the $\mathbf{F}_a$'s is orthogonal to one of the $\mathbf{E}_a$'s:

$$\int_V \mathbf{F}_a\cdot\mathbf{E}_b dv=0. \quad \text{(III.20)}$$

To prove this, we note that

$$\text{div}\,(\psi_a\mathbf{E}_b)=\psi_a\,\text{div}\,\mathbf{E}_b+\text{grad}\,\psi_a\cdot\mathbf{E}_b =k_a\mathbf{F}_a\cdot\mathbf{E}_b, \quad \text{(III.21)}$$

since $\text{div}\,\mathbf{E}_b=0$. Integrating over V, the surface integral vanishes, proving our result (III.20).

We have now set up two orthogonal families of solenoidal functions, the $\mathbf{E}_a$'s and $\mathbf{H}_a$'s, and one family of irrotational functions, the $\mathbf{F}_a$'s. It seems intuitively reasonable to suppose that the $\mathbf{E}_a$'s and the $\mathbf{F}_a$'s, or the $\mathbf{H}_a$'s and the $\mathbf{F}_a$'s, form complete sets of functions, in the mathematical sense, such that any artibrary vector function of position within V, satisfying certain not very stringent conditions of continuity, must be capable of being expanded in a series in the functions. It is a problem for the mathematician to prove rigorously the completeness of these sets of functions, and we shall not attempt it. There are two points which are clear about this expansion. In the first place, we certainly do not need both the $\mathbf{E}_a$'s and the $\mathbf{H}_a$'s for any given expansion; an $\mathbf{E}_a$ and an $\mathbf{H}_b$ are not orthogonal to each other, but on the contrary one of the $\mathbf{E}$'s can be expanded in series in the $\mathbf{H}$'s, or *vice versa*. We choose either the $\mathbf{E}_a$'s or the $\mathbf{H}_a$'s, in any given case, according to convenience. The general situation to be expected is that if the function which we wish to expand has boundary conditions which more closely resemble the boundary conditions satisfied by the $\mathbf{E}_a$'s than by the $\mathbf{H}_a$'s, then its expansion in terms of the $\mathbf{E}_a$'s will converge better than the expansion in terms of the $\mathbf{H}_a$'s, and *vice versa*. This is analogous to the situation according to which a function in the range 0 to π can be expanded either in a sine or a cosine series of period 2π, but if the function we are expanding satisfies the same boundary conditions as the sines (function zero at 0 and π), the expansion in sines will converge better than that in cosines, in the sense that the derivative of the sine series will also converge, while the derivative of the cosine series will diverge. The other point which we wish to make about the expansion is that the characteristic numbers k_a for the solenoidal functions are not the same as for the irrotational functions. In each case, for instance in the summations over a which we shall soon set up, we are to understand that we use the k_a appropriate for the type of function under consideration.

If we may expand an arbitrary function, say $\mathbf{A}$, in terms of the $\mathbf{E}_a$'s and $\mathbf{F}_a$'s (or the $\mathbf{H}_a$'s and $\mathbf{F}_a$'s), it is then an easy matter from the orthogonality and normalization conditions to find the expansion coefficients. Let us assume that

$$\mathbf{A}=\mathbf{A}_1+\mathbf{A}_2, \quad \text{(III.22)}$$

where $\mathbf{A}_1$ is the solenoidal, $\mathbf{A}_2$ the irrotational, part of $\mathbf{A}$. Then $\mathbf{A}_1$ can be expanded in series in the $\mathbf{E}_a$'s (or the $\mathbf{H}_a$'s), and $\mathbf{A}_2$ in the $\mathbf{F}_a$'s:

$$\mathbf{A}=\textstyle\sum_a e_a\mathbf{E}_a+\sum_a f_a\mathbf{F}_a, \quad \text{(III.23)}$$

where the e_a's and f_a's are coefficients, and where the first summation is over the normal functions of the solenoidal, the second of the irrotational, type. If we multiply $\mathbf{A}$ by one of the $\mathbf{E}_a$'s, and integrate over V, then on account of the orthogonality relations the integral of the product of this $\mathbf{E}_a$ with every other term of the summation (III.23) except itself will be zero, and the integral of its square will be unity on account of normalization. Thus we have

$$e_a=\int_V \mathbf{A}\cdot\mathbf{E}_a dv. \quad \text{(III.24)}$$

There will be a similar result for f_a. Hence we may rewrite (III.23) in the form

$$\mathbf{A}=\sum_a \mathbf{E}_a \int_V \mathbf{A}\cdot\mathbf{E}_a dv+\mathbf{F}_a \int_V \mathbf{A}\cdot\mathbf{F}_a dv. \quad \text{(III.25)}$$

There will be an analogous formula in terms of the $\mathbf{H}_a$'s. Our method of expansion, which we take up in the next section, is based on this result. In expressing the volume integrals in the future, we shall omit the subscript V for simplicity, assuming that all volume integrals are over V unless it is stated to the contrary.

2. Maxwell's Equations in a Hollow Cavity

Our object in the present section will be to expand the electric and magnetic fields, and other related quantities, inside a hollow cavity in terms of the orthogonal functions set up in the preceding section, and by means of Maxwell's equations to find relations between the expansion coefficients. In Maxwell's equations, as given in (II.2) and (II.3), we shall expand $\mathbf{E}$ in series in the $\mathbf{E}_a$'s and $\mathbf{F}_a$'s; $\mathbf{H}$ in series in the $\mathbf{H}_a$'s; curl $\mathbf{E}$ in the $\mathbf{H}_a$'s; curl $\mathbf{H}$ in the $\mathbf{E}_a$'s; $\mathbf{J}$ in the $\mathbf{E}_a$'s and $\mathbf{F}_a$'s; div $\mathbf{D}$ and ρ in the ψ_a's. Thus we have

$$\begin{aligned}
\mathbf{E}&=\sum_a\left(\mathbf{E}_a\int\mathbf{E}\cdot\mathbf{E}_a dv+\mathbf{F}_a\int\mathbf{E}\cdot\mathbf{F}_a dv\right),\\
\mathbf{H}&=\sum_a \mathbf{H}_a\int\mathbf{H}\cdot\mathbf{H}_a dv,\\
\mathbf{J}&=\sum_a\left(\mathbf{E}_a\int\mathbf{J}\cdot\mathbf{E}_a dv+\mathbf{F}_a\int\mathbf{J}\cdot\mathbf{F}_a dv\right),\\
\rho&=\sum_a \psi_a\int\rho\psi_a dv.
\end{aligned} \quad \text{(III.26)}$$

The other expansions take a little more thought. For curl $\mathbf{E}$, we might think at first sight that we could take the expansion (III.26) for $\mathbf{E}$, and take its curl directly, obtaining $\sum k_a\mathbf{H}_a\int\mathbf{E}\cdot\mathbf{E}_a dv$. On the other hand, we might consider the function curl $\mathbf{E}$, and expand it directly in terms of the functions $\mathbf{H}_a$. We then have

$$\text{curl }\mathbf{E}=\sum_a \mathbf{H}_a\int \text{curl }\mathbf{E}\cdot\mathbf{H}_a dv. \quad \text{(III.27)}$$

We may evaluate the integral by the following method. We have

$$\begin{aligned}
\text{div }(\mathbf{E}\times\text{curl }\mathbf{E}_a)&=\text{curl }\mathbf{E}\cdot\text{curl }\mathbf{E}_a-\mathbf{E}\cdot\text{curl curl }\mathbf{E}_a\\
&=k_a\mathbf{H}_a\cdot\text{curl }\mathbf{E}-k_a^2\mathbf{E}\cdot\mathbf{E}_a.
\end{aligned} \quad \text{(III.28)}$$

Integrating over V, and transforming the term on the left into a surface integral,

$$\int_{S,S'}\mathbf{n}\cdot(\mathbf{E}\times\mathbf{H}_a)da=\int\text{curl }\mathbf{E}\cdot\mathbf{H}_a dv-k_a\int\mathbf{E}\cdot\mathbf{E}_a dv. \quad \text{(III.29)}$$

In the surface integral, we may write the integrand in the form $\mathbf{H}_a\cdot(\mathbf{n}\times\mathbf{E})$ or $\mathbf{E}\cdot(\mathbf{H}_a\times\mathbf{n})$. The second form shows that it vanishes over S'. Then, substituting in (III.27), we have

$$\text{curl }\mathbf{E}=\sum_a \mathbf{H}_a\left(k_a\int\mathbf{E}\cdot\mathbf{E}_a dv+\int_S(\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a da\right). \quad \text{(III.30)}$$

In other words, in addition to the sum of terms involving volume integrals, which we got by taking the curl of the series for $\mathbf{E}$, we also have a sum of terms involving surface integrals. These integrals will be zero if $\mathbf{E}$ is normal to the surface S, but we shall later find cases where $\mathbf{E}$ has a tangential component over S, so that the surface integral does not vanish. We shall shortly consider the significance of these surface integrals. In a corresponding way we find that

$$\text{curl }\mathbf{H}=\sum_a \mathbf{E}_a\left(k_a\int\mathbf{H}\cdot\mathbf{H}_a dv+\int_{S'}(\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da\right). \quad \text{(III.31)}$$

Similarly for div $\mathbf{D}$, using the relation

$$\begin{aligned}
\text{div }(\psi_a\mathbf{D})&=\psi_a\text{ div }\mathbf{D}+\mathbf{D}\cdot\text{grad }\psi_a\\
&=\psi_a\text{ div }\mathbf{D}+k_a\mathbf{D}\cdot\mathbf{F}_a,
\end{aligned} \quad \text{(III.32)}$$

we have

$$\text{div }\mathbf{D}=\sum_a \psi_a\left(-k_a\int\mathbf{D}\cdot\mathbf{F}_a dv+\int_{S',S}(\mathbf{D}\cdot\mathbf{n})\psi_a da\right). \quad \text{(III.33)}$$

The surface integral in (III.33) vanishes because ψ_a is zero on S and S'.

We have now set up the series for the various quantities concerned in Maxwell's equations. We next substitute these series, and equate coefficients, so as to get the differential equations satisfied by the various coefficients. From curl $\mathbf{E} + \partial\mathbf{B}/\partial t = 0$ we have

$$k_a\int \mathbf{E}\cdot\mathbf{E}_a dv + \mu_0\frac{d}{dt}\int \mathbf{H}\cdot\mathbf{H}_a dv = -\int_S (\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a da. \quad \text{(III.34)}$$

From curl $\mathbf{H} - \dot{\mathbf{D}} = \mathbf{J}$ we have

$$k_a\int \mathbf{H}\cdot\mathbf{H}_a dv - \epsilon_0\frac{d}{dt}\int \mathbf{E}\cdot\mathbf{E}_a dv = \int \mathbf{J}\cdot\mathbf{E}_a dv - \int_{S'} (\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da, \quad \text{(III.35)}$$

$$-\epsilon_0\frac{d}{dt}\int \mathbf{E}\cdot\mathbf{F}_a dv = \int \mathbf{J}\cdot\mathbf{F}_a dv. \quad \text{(III.36)}$$

The equation div $\mathbf{B} = 0$ is automatically satisfied. The equation div $\mathbf{D} = \rho$ gives

$$-k_a\epsilon_0\int \mathbf{E}\cdot\mathbf{F}_a dv = \int \rho\psi_a dv. \quad \text{(III.37)}$$

We may show from the equation of continuity that (III.36) and (III.37) are equivalent. Taking the time derivative of (III.37), and multiplying (III.36) by k_a, the left sides of the equations are identical. The two equations will then lead to the same result if we have

$$k_a\int \mathbf{J}\cdot\mathbf{F}_a dv = \frac{d}{dt}\int \rho\psi_a dv. \quad \text{(III.38)}$$

The integral on the left may be rewritten by using the relation

$$\text{div}\,(\psi_a\mathbf{J}) = \psi_a\,\text{div}\,\mathbf{J} + \mathbf{J}\cdot\text{grad}\,\psi_a. \quad \text{(III.39)}$$

Integrating (III.39) over V, the surface integral is zero, and the volume integral, substituted in (III.38), at once leads to

$$-\int \text{div}\,\mathbf{J}\psi_a dv = \frac{d}{dt}\int \rho\psi_a dv. \quad \text{(III.40)}$$

But (III.40) is true on account of the equation of continuity,

$$\text{div}\,\mathbf{J} + \partial\rho/\partial t = 0. \quad \text{(III.41)}$$

Thus we need use only (III.36) or (III.37), not both.

In discussing Maxwell's equations, we have two very distinct problems. First there are the Eqs. (III.34) and (III.35). These will concern us most in the future. They are the equations determining the solenoidal part of $\mathbf{E}$, and $\mathbf{H}$. It is this part of the field that shows properties of wave propagation, and that is usually regarded as the radiation field. These equations determine the coefficients $\int \mathbf{E}\cdot\mathbf{E}_a dv$ and $\int \mathbf{H}\cdot\mathbf{H}_a dv$ as functions of time; they are determined in terms of the integrals which appear on the right side of the equations. The quantity $\int \mathbf{J}\cdot\mathbf{E}_a dv - \int_{S'} (\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da$ appearing on the right side of (III.35) is easy to interpret. It is one of the components of $\mathbf{J}$ (the volume integral), supplemented by the surface integral. In this surface integral $-(\mathbf{n}\times\mathbf{H})$ is the tangential component of surface current connected with the discontinuity in the tangential component of $\mathbf{H}$ at the surface, so that the surface integral gives us the contribution of surface currents. Similarly the quantity $-\int_S (\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a da$ would have to be interpreted as the integral of $-(\mathbf{n}\times\mathbf{E})$, a surface density of fictitious magnetic current, which appears at a surface of discontinuity of the tangential component of $\mathbf{E}$. Thus we have the interpretation of the surface integrals appearing in (III.30) and (III.31).

We can easily combine (III.34) and (III.35), to get separate equations for $\int \mathbf{E}\cdot\mathbf{E}_a dv$, $\int \mathbf{H}\cdot\mathbf{H}_a dv$, as in the conventional derivation of the wave equation. Thus we have

$$\epsilon_0\mu_0\frac{d^2}{dt^2}\int \mathbf{E}\cdot\mathbf{E}_a dv + k_a^2\int \mathbf{E}\cdot\mathbf{E}_a dv = -\mu_0\frac{d}{dt}\left(\int \mathbf{J}\cdot\mathbf{E}_a dv - \int_{S'} (\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da\right) - k_a\int_S (\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a da, \quad \text{(III.42)}$$

$$\epsilon_0\mu_0\frac{d^2}{dt^2}\int \mathbf{H}\cdot\mathbf{H}_a dv + k_a^2\int \mathbf{H}\cdot\mathbf{H}_a dv$$
$$= k_a\left(\int \mathbf{J}\cdot\mathbf{E}_a dv - \int_{S'}(\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da\right)$$
$$-\epsilon_0\frac{d}{dt}\int_S(\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a da. \quad \text{(III.43)}$$

These two equations will form the basis of most of our later treatment of resonant cavities. They show that $\int\mathbf{E}\cdot\mathbf{E}_a dv$ and $\int\mathbf{H}\cdot\mathbf{H}_a dv$ are determined as functions of time by the type of differential equations encountered in elementary problems of simple harmonic motion. The terms on the right-hand sides take the place of the external force in the problem of simple harmonic motion, so that we may get solutions of (III.42) and (III.43) showing the properties of forced oscillations and resonance, the forced motion resulting from the currents within the cavity, or from disturbances propagated from the walls. We shall take up these motions in later sections.

The remaining one of Eqs. (III.36) or (III.37) refers to the irrotational part of **E**, the part whose curl is zero, or which is derivable from a scalar potential. To understand its interpretation, let us go back to Maxwell's equations, writing the vectors **E** and **J** as sums of a solenoidal part ($\mathbf{E}_1$ and $\mathbf{J}_1$) and an irrotational part ($\mathbf{E}_2$ and $\mathbf{J}_2$). We have

$$\text{curl } \mathbf{E}_1+\dot{\mathbf{B}}=0, \quad \text{div } \mathbf{B}=0,$$
$$\text{curl } \mathbf{H}-\dot{\mathbf{D}}_1=\mathbf{J}_1, \quad \text{div } \mathbf{D}_1=0,$$
$$-\dot{\mathbf{D}}_2=\mathbf{J}_2, \quad \text{div } \mathbf{D}_2=\rho$$

in which the last two equations are equivalent as a result of the equation of continuity. Remembering that curl $\mathbf{E}_2=0$, we may write

$$\mathbf{E}_2=-\text{grad } \phi, \quad \nabla^2\phi=-\rho/\epsilon_0. \quad \text{(III.44)}$$

That is, $\mathbf{E}_2$ is derivable from a scalar potential, which satisfies Poisson's equation. The problem of finding $\mathbf{E}_2$ is then identical with the electrostatic problem of finding the field of a known distribution of charge, subject to a condition that the potential is zero on the boundary. The only difference between this and an electrostatic problem is that the charge distribution, and hence the field, varies with time; but no account is taken in the problem of retardation, or of finite velocity of propagation of the disturbance.

3. Free and Damped Oscillations of a Resonant Cavity

As a first example of the use of Eqs. (III.42) and (III.43), let us consider the free oscillations of a resonant cavity. We assume that the cavity contains no current density, so that $\mathbf{J}=0$. Furthermore we assume that over the part of the surface S the tangential component of **E** is zero (which would be the case if that part of the surface were formed of a perfect conductor, or formed a short circuit) and that over S' the tangential component of **H** is zero (which would be the case if that part of the surface were formed of a perfect insulator, in which no surface current could flow, so that it formed an open circuit). Then all integrals on the right side of (III.42) and (III.43) would be zero, and the equations would have solutions

$$\int\mathbf{E}\cdot\mathbf{E}_a dv=\text{constant } e^{j\omega_a t}, \quad \omega_a^2\epsilon_0\mu_0=k_a^2 \quad \text{(III.45)}$$

with similar solutions for $\int\mathbf{H}\cdot\mathbf{H}_a dv$. Thus the ω_a's are the angular frequencies of the resonant modes, and the general solution of the problem of free oscillations would be a superposition of the various normal modes, each oscillating with arbitrary amplitude at its resonant frequency. From Maxwell's Eqs. (III.34) and (III.35) we can find the relation between the coefficients $\int\mathbf{E}\cdot\mathbf{E}_a dv$ and $\int\mathbf{H}\cdot\mathbf{H}_a dv$ for the normal modes. By substituting, and using (III.45), we find easily that

$$\frac{\int\mathbf{E}\cdot\mathbf{E}_a dv}{\int\mathbf{H}\cdot\mathbf{H}_a dv}=-j(\mu_0/\epsilon_0)^{\frac{1}{2}}. \quad \text{(III.46)}$$

That is, the magnitudes of the coefficients $\int\mathbf{E}\cdot\mathbf{E}_a dv$ and $\int\mathbf{H}\cdot\mathbf{H}_a dv$ are in the same ratios to each other as the values of **E** and **H** in a plane wave in empty space, but the electric and magnetic fields are 90° apart in phase, a characteristic of standing electromagnetic waves. We readily find that as a result of (III.46) the time average electrical energy, integrated through the cavity, equals the time average magnetic energy. The phase difference, however, results in the magnetic energy being large when the electric energy is small, and *vice versa*, just as with the kinetic and potential energy in simple harmonic motion, with the result that the total energy remains constant.

The solution for free oscillation which we have just found is analogous to the free oscillation of a simple L-C series circuit. We shall now look for the analog to damped oscillation, which occurs when the circuit contains resistance as well as inductance and capacity. The equation for a series circuit containing inductance L, resistance R, and capacity C is

$$L\frac{d^2q}{dt}+R\frac{dq}{dt}+\frac{q}{C}=0, \qquad \text{(III.47)}$$

if we use the charge q on the condenser as the variable. If we assume an exponential solution, q varying as $e^{j\omega t}$, this becomes

$$\left(-L\omega^2+Rj\omega+\frac{1}{C}\right)q=0,$$

which may be rewritten in the form

$$\frac{\omega}{\omega_0}-\frac{\omega_0}{\omega}-j\frac{R}{L\omega_0}=0, \quad \text{where} \quad \omega_0{}^2=1/LC. \qquad \text{(III.48)}$$

Equation (III.48) is a quadratic for the frequency, whose solution is found to be

$$\frac{\omega}{\omega_0}=\frac{j}{2Q}\pm[1-(1/2Q)^2]^{\frac{1}{2}},$$

where

$$\frac{1}{Q}=\frac{R}{L\omega_0}. \qquad \text{(III.49)}$$

Using the value (III.49) for the frequency, we see that the charge varies as

$$\exp\left(-(\omega_0/2Q)t\right)\exp\left(\pm j\omega_0[1-(1/2Q)^2]^{\frac{1}{2}}t\right). \qquad \text{(III.50)}$$

This represents a damped oscillation, with angular frequency equal to $\omega_0[1-(1/2Q)^2]^{\frac{1}{2}}$, and such that the energy, which is proportional to the square of the amplitude, decreases with time according to the exponential function $e^{-(\omega_0/Q)t}$. The decrease in energy per unit time is ω_0/Q times the energy, or we may write

$$Q=\frac{2\pi\times\text{total energy}}{\text{decrease of energy per period}}, \qquad \text{(III.51)}$$

where the period concerned is $2\pi/\omega_0$, as determined from the angular frequency which would exist if the damping were absent.

The description of a rate of damping of a circuit by means of a Q is one which is convenient in microwave work as well as with ordinary oscillating circuits. We shall adopt for Q a definition which is equivalent to (III.48): an oscillation whose angular frequency is determined by the equation

$$j\left(\frac{\omega}{\omega_0}-\frac{\omega_0}{\omega}\right)+\frac{1}{Q}=0 \qquad \text{(III.52)}$$

will be referred to as having a given Q. This definition is equivalent to that of (III.51). In most of our applications, Q will be large enough so that the distinction between ω_0 and the corrected angular frequency $\omega_0[1-(1/2Q)^2]^{\frac{1}{2}}$ can be neglected. The advantage of introducing Q and ω_0 is that Q and ω/ω_0 are dimensionless quantities, easily measured, and easy to transfer to microwave problems in which L, R, and C have only a rather uncertain or ambiguous significance. In many cases we shall find that the angular frequency ω_0 is nearly equal to one of the resonant frequencies ω_a defined in (III.45); this will be the case if the motion differs from a free oscillation only by a small perturbation. In such a case, we may let

$$\omega_0=\omega_a+\Delta\omega_a \qquad \text{(III.53)}$$

where $\Delta\omega_a$ is a small quantity. In this case, (III.52) may be rewritten in several forms, correct to the first order of small quantities, as follows:

$$j\left(\frac{\omega}{\omega_a}-\frac{\omega_a}{\omega}\right)+\frac{1}{Q}-2j\frac{\Delta w_a}{\omega_a}=0,$$

$$\omega^2-\omega_a{}^2=j\frac{\omega_a{}^2}{Q}+2\omega_a\Delta\omega_a,$$

$$\omega=\omega_a+\Delta\omega_a+j\frac{\omega_a}{2Q}. \qquad \text{(III.54)}$$

We see from (III.54) that the real part of the quantity $\omega^2-\omega_a^2$ leads to a shift of resonant frequency, and the imaginary part to a damping of the oscillation.

We now ask what sort of perturbation of our problem of the free oscillation of a cavity is necessary to produce damped oscillations. Three types of perturbation are most common. First, the walls of the cavity, instead of being perfect conductors, may have only finite conductivity, resulting in resistive losses. In this case, the tangential component of $\mathbf{E}$ over the surface S will not vanish, and the integral $\int_S (\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a$ in (III.42) will be different from zero. Secondly, the cavity may have certain windows, or wave guide outputs, which allow the escape of energy, with consequent decrease of the total energy. In this case, the tangential component of $\mathbf{H}$ over the surface S' will not be zero, and the integral $\int_{S'} (\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da$ will not vanish. Thirdly, the cavity may contain currents in phase with the voltages producing them, as if they contained resistive material obeying Ohm's law. In that case, the integral $\int \mathbf{J}\cdot\mathbf{E}_a dv$ will not vanish. In all these cases, in order to have damped oscillations, the various integrals must be proportional to the amplitude $\int \mathbf{E}\cdot\mathbf{E}_a dv$ of the field component, and must oscillate with the same period. In such a case, we may assume a solution of (III.42) varying as $e^{j\omega t}$, where ω will in general be complex. Substituting this time variation, and using (III.54), we find

$$\frac{1}{Q}-2j\frac{\Delta\omega_a}{\omega_a}=\frac{-j}{\omega_a(\epsilon_0\mu_0)^{\frac{1}{2}}}\frac{\int_S (\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a da}{\int \mathbf{E}\cdot\mathbf{E}_a dv}-\frac{1}{\epsilon_0\omega_a}\frac{\int_{S'} (\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da}{\int \mathbf{E}\cdot\mathbf{E}_a dv}+\frac{1}{\epsilon_0\omega_a}\frac{\int \mathbf{J}\cdot\mathbf{E}_a dv}{\int \mathbf{E}\cdot\mathbf{E}_a dv}. \quad \text{(III.55)}$$

We shall now consider the various terms on the right-hand side, showing how they lead to damping and to displacement of the resonant frequency.

4. The Unloaded Q

First we consider the losses in the cavity resulting from the finite conductivity of the walls. In (II.34) we have found that there is a tangential component of $\mathbf{E}$ over the surface S, in case of finite conductivity. From that equation we find at once that

$$\mathbf{n}\times\mathbf{E}=\mathbf{H}(\mu_0\omega/2\sigma)^{\frac{1}{2}}(1+j)=\mathbf{H}\tfrac{1}{2}\delta\mu_0\omega(1+j) \quad \text{(III.56)}$$

at the surface. Thus the first term on the right side of (III.55) becomes

$$\frac{-j}{\omega_a(\epsilon_0\mu_0)^{\frac{1}{2}}}\frac{\int_S (\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a da}{\int \mathbf{E}\cdot\mathbf{E}_a dv}=-j(\mu_0/\epsilon_0)^{\frac{1}{2}}(1+j)\frac{\int_S \frac{1}{2}\delta\mathbf{H}\cdot\mathbf{H}_a da}{\int \mathbf{E}\cdot\mathbf{E}_a dv}. \quad \text{(III.57)}$$

If the oscillation is taking place with almost the frequency of the ath mode, we may assume that to a first approximation the field distribution will be that of the ath mode as well. Thus we shall approximately have

$$\mathbf{H}=\mathbf{H}_a\int \mathbf{H}\cdot\mathbf{H}_a dv=j(\epsilon_0/\mu_0)^{\frac{1}{2}}\mathbf{H}_a\int \mathbf{E}\cdot\mathbf{E}_a dv \quad \text{(III.58)}$$

in which we have used (III.46) for the ratio between $\int \mathbf{E}\cdot\mathbf{E}_a dv$ and $\int \mathbf{H}\cdot\mathbf{H}_a dv$, a value which strictly holds only for the free oscillation, but which would be a good approximation for damped oscillation. Substituting, we have

$$1/Q-2j(\Delta\omega_a/\omega_a)=(1+j)\int_S \tfrac{1}{2}\delta H_a^2 da. \quad \text{(III.59)}$$

We see from (III.59) that the surface losses result in a shift of wave-length, as well as a contribution to Q. The value of Q given in (III.59) is generally called the unloaded Q; we shall denote it by Q_a. We note that it really should also have a subscript a to denote the ath mode, but we shall omit this when it is not necessary. We may get an idea of the order of magnitude of the unloaded Q as follows. We remember that from the normalization condition (III.11) the integral $\int H_a^2 dv$ is equal to unity. That is, if V is the volume of the cavity, and if $\langle H_a^2 \rangle_{\text{Av}}$ is the mean value of H_a^2, we have $\langle H_a^2 \rangle_{\text{Av}} = 1/V$. If we assume, to get orders of magnitude, that δ is constant over the surface, and that the value of H_a^2 on the surface equals its average over the volume and that furthermore the surface area is A, then we should have $1/Q_a = \delta A/2V$. That is, Q_a would be the ratio of the volume, to the volume of a thin shell of thickness $\delta/2$ surrounding the volume. It is interesting to see how the Q_a of a cavity will change with the wave-length. Of two cavities of the same shape but different sizes, the wave-length will be proportional to the linear dimensions, so that the volume will vary as λ^3, and the area as λ^2. The skin depth, by (II.33), is proportional to $(\lambda)^{\frac{1}{2}}$. Thus if the conductivity is independent of wave-length, Q_a will be proportional to $(\lambda)^{\frac{1}{2}}$, decreasing as we go to shorter wave-lengths. The actual magnitude of Q_a will of course depend on the shape of the cavity, and the material of which it is made.

5. The Input Impedance of a Cavity

We next take up the effect of coupling the cavity to an outside system by an output lead, which will be assumed to take the form of a wave guide or coaxial line. We assume that there is such a line attached to the cavity, and that the surface S' is a surface at a cross section of the line. The volume in which we are solving Maxwell's equations then includes not merely the cavity, but the part of the output lead out to the surface S'. The normal resonant mode is that which arises when there is an open circuit at S'; that is, when there is an infinite standing wave ratio in the output line, with a standing wave maximum at S'. This would correspond to a voltage maximum, and current node, at this surface. Our problem is now to substitute other boundary conditions at S', and to find what effect that has on the oscillations in the cavity. If the boundary condition consists of stating that the impedance or admittance across S' has a certain definite value, so that the voltage is proportional to the current (or $\mathbf{E}$ is proportional to $\mathbf{H}$), the result will be a contribution to Q and the frequency shift. The more general case, however, is one in which there are arbitrary impressed voltages or currents across S'. In such a case, power can flow in as well as out through the guide, so that we have the possibility of forced oscillation of the cavity as well as damped oscillation. We take up this general case, later obtaining the case of damping as a special case of our general treatment.

The general outline of our derivation will be as follows. We assume a given distribution of $\mathbf{H}$ over the surface S', or a given current flowing in the guide. We can then calculate the integral $\int(\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da$ over S'. From (III.42) we can compute the electric field everywhere within the cavity, and in particular within the guide, and at the surface S'. From this electric field we can find the voltage at S', and can take the ratio of voltage to current, and hence the impedance at the plane S', the input impedance looking into the cavity. First we must consider the nature of the field in the wave guide. In the guide, the function H_a must have the general form given by a summation over the modes n of the guide of terms as given in (II.13); for any solution of Maxwell's equations in the guide must have that form. Thus at the surface S' the transverse component of E_a must be expressible in the form

$$\mathbf{E}_a = \sum_n \mathbf{E}_{tn}(v_{an}/Z_{0n}) \qquad \text{(III.60)}$$

where Z_{0n} is the characteristic impedance of the guide in the nth mode, for an angular frequency ω_a, and where the v_{an}'s are coefficients (independent of time, as $\mathbf{E}_a$ is) which as we see from (II.22) represent the voltage at plane S' and in the nth wave guide mode set up by the resonance $\mathbf{E}_a$ in the cavity. The transverse component of $\mathbf{H}_a$ is zero at S', by hypothesis; that is, as we mentioned earlier, the corresponding current components are zero at this plane. If the transverse component of $\mathbf{H}_a$ is zero, then by (II.13) the normal component of $\mathbf{E}_a$ will automatically be zero on S'. It is interesting to consider the effect on the v_{an}'s of choosing the surface S' at different points along the output line. If we go a half guide wave-length along the guide, for any of the propagated modes, the disturbance will come back to its initial value. Thus v_{an}, though it will vary with the position of S', will be a periodic function. We shall discuss the implications of this periodicity at a later point. For the attenuated modes, on the contrary, the disturbance will generally fall off exponentially as we go along the guide away from the cavity (there is normally no reason to expect the other exponential term, which increases exponentially as we go away from the cavity, to be present). Thus if we take S' some little distance away from the cavity, the values of the v_{an}'s for the attenuated waves will have fallen to very low values, and may be neglected. As a rule we shall assume that our surface S' is far enough from the cavity for us to have this situation. Then our summation over n really has non-vanishing terms only for the propagated waves. In case we are considering an angular frequency ω_a for which only the dominant mode in the guide will be propagated, the summation will reduce to a single term.

We shall now assume that over the surface S' we impose a tangential magnetic field which in accordance with (II.13) and (II.22) we may write

$$\mathbf{H} = \sum_n \mathbf{H}_{tn} i_n \qquad \text{(III.61)}$$

where the i_n's can be interpreted, as in Chapter II, as the currents associated with the various modes of the guide. We assume that $\mathbf{H}$ varies as $e^{j\omega t}$, each of the i_n's having this variation. We may now compute the integral over S' of $(\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a$, which appears on the right side of (III.42). We shall take the positive direction in the guide (the direction of the unit vector $\mathbf{k}$ of Chapter II) as being into the cavity, so that the unit vector $\mathbf{n}$ will be $-\mathbf{k}$. From (II.11), we then have $\mathbf{n}\times\mathbf{H} = \sum_n i_n \mathbf{E}_{tn}/Z_{0n}$. Multiplying this by the value of $\mathbf{E}_a$ from (III.60), and integrating over the surface S', the cross terms in the double sum integrate to zero on account of the orthogonality properties of Section 5, Chapter II, and the result, using the normalization condition (II.30), is

$$\int_{S'} (\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da = \sum_n i_n v_{an}. \qquad \text{(III.62)}$$

If now we use this value of the integral in (III.42), and assume that $\int\mathbf{E}\cdot\mathbf{E}_a dv$ varies as $e^{j\omega t}$, we find at once that

$$\int \mathbf{E}\cdot\mathbf{E}_a dv = \sum_n \frac{i_n v_{an}/\epsilon_0\omega_a}{j[(\omega/\omega_a)-(\omega_a/\omega)]}. \qquad \text{(III.63)}$$

The voltage corresponding to the nth guide mode, if $\mathbf{E}=\mathbf{E}_a$, or if $\int\mathbf{E}\cdot\mathbf{E}_a dv = 1$, is v_{an}. Thus the whole voltage corresponding to the nth guide mode is $\sum_a v_{an}\int\mathbf{E}\cdot\mathbf{E}_a dv$. If we call this voltage V_n, we have

$$V_n = \sum_m i_m Z_{nm},$$
$$Z_{nm} = \sum_a \frac{v_{an}v_{am}/\epsilon_0\omega_a}{j[(\omega/\omega_a)-(\omega_a/\omega)]}. \qquad \text{(III.64)}$$

That is, there are linear relationships between the V_n's and the i_m's, and we have been able to write down the coefficients, of the nature of impedances, in an explicit manner.

There are a number of remarks which can be made about this important result. In the first place, suppose that we are applying it in a region of wave-lengths where the guide can propagate only in its dominant mode. Then in the summation over m, we shall have only one term, that related to the dominant mode; all the other Z_{nm}'s will vanish. If this mode is called the 1st mode, we shall have

$$V_1 = i_1 Z_{11},$$
$$\frac{V_1}{i_1} = Z_{11} = \sum_a \frac{v_{a1}^2/\epsilon_0\omega_a}{j[(\omega/\omega_a)-(\omega_a/\omega)]}. \qquad \text{(III.65)}$$

This then represents the impedance seen looking into a cavity through a wave guide which propagates only in the dominant mode. We observe that it is a sum of resonant terms, the impedance becoming infinite when the frequency equals the resonant frequency of any one of the modes. At frequencies near ω_a, the term in this frequency is much greater than any other, and varies rapidly with frequency, whereas the other terms are small and slowly varying. It is often convenient to lump these other terms together, using a notation Z_{a1} to represent them, so that

$$Z_{11}=\frac{v_{a1}^2/\epsilon_0\omega_a}{j[(\omega/\omega_a)-(\omega_a/\omega)]}+Z_{a1}. \quad \text{(III.66)}$$

We may now ask what will be the damping if the guide is terminated by an impedance Z_1. The quantity Z_{11} in (III.66) is the impedance looking into the cavity, or is the negative of Z_1, the impedance looking out. Making this substitution, we have from (III.66)

$$j\left(\frac{\omega}{\omega_a}-\frac{\omega_a}{\omega}\right)+\frac{v_{a1}^2/\epsilon_0\omega_a}{Z_1+Z_{a1}}=0. \quad \text{(III.67)}$$

This is in the form of the first of Eqs. (III.54), and shows that we have

$$\frac{1}{Q}-2j\frac{\Delta\omega_a}{\omega_a}=\frac{v_{a1}^2/\epsilon_0\omega_a}{Z_1+Z_{a1}}=\frac{1}{Q_{\text{ext},a1}}\frac{Z_{01}}{Z_1+Z_{a1}},$$

where

$$\frac{1}{Q_{\text{ext},a1}}=\frac{v_{a1}^2}{\epsilon_0\omega_a Z_{01}}. \quad \text{(III.68)}$$

In this formula, we have introduced a quantity, $Q_{\text{ext},a1}$, which we may call the external Q of the ath mode of the cavity, and the 1st mode of the guide. Its meaning is simple; if we make $Z_1+Z_{a1}=Z_{01}$, it is the Q of the resonant mode. If we neglect Z_{a1}, this means that there is to be a matched load in the guide. We shall see presently that if we choose the position of S' properly, we can make Z_{a1} equal to zero, so that in this case Q_{ext} is exactly the Q which we should have with a matched load. The external Q is clearly a measure of the coupling of the ath mode to the output, through the 1st mode of the guide. In case v_{a1} is very small, the external Q is large, or there is very small coupling. It is convenient to define an admittance $g+jb$ by the relation

$$g+jb=Z_{01}/(Z_1+Z_{a1}). \quad \text{(III.69)}$$

In this case, we have

$$\frac{1}{Q}=\frac{g}{Q_{\text{ext},a1}},\qquad -2\frac{\Delta\omega_a}{\omega_a}=\frac{b}{Q_{\text{ext},a1}}. \quad \text{(III.70)}$$

In terms of the external Q, we may rewrite (III.65) in the form

$$\frac{Z_{11}}{Z_{01}}=\sum_a\frac{1/Q_{\text{ext},a1}}{j[(\omega/\omega_a)-(\omega_a/\omega)]}$$
$$=\frac{1/Q_{\text{ext},a1}}{j[(\omega/\omega_a)-(\omega_a/\omega)]}+\frac{Z_{a1}}{Z_{01}}. \quad \text{(III.71)}$$

In the treatment we have just given, we have neglected the other terms on the right side of (III.42), coming from losses in the walls and other forms of losses. If however we have the situation of Section 4, we may easily take account of the losses in the walls. When the external frequency is near the resonant frequency of the ath mode, the term in a in the summation (III.65) will be much greater than any other, which means that the field is almost like $\mathbf{E}_a$, the assumption underlying Section 4. In setting up (III.63), we then see that we must put additional terms as given by (III.59) in the denominator. We then have in place of (III.71)

$$\frac{Z_{11}}{Z_{01}}=\sum_a\frac{1/Q_{\text{ext},a1}}{j[(\omega/\omega_a)-(\omega_a'/\omega)]+(1/Q_a)}, \quad \text{(III.72)}$$

where we have used ω_a' to refer to the resonant frequency as modified by the correction term $\Delta\omega_a$ derived from (III.59). We see that the input impedance is no longer purely reactive, but that it has a resistive term, the resonance term becoming purely resistive at resonance, just like the input impedance of a parallel resonant circuit in ordinary circuit theory. Proceeding as in the derivation of (III.68), we now find that if there is loss in the walls, the quantity $1/Q-2j\Delta\omega_a/\omega_a$ is the sum of the quantities (III.59) and (III.68) arising from the losses in the walls, and from the effect of the output lead. The resulting Q is called the loaded Q:

$$\frac{1}{Q_L}=\frac{1}{Q_a}+\frac{g}{Q_{\text{ext},a1}}. \quad \text{(III.73)}$$

We next consider the case where there are several output leads for the cavity. In (III.62), the integration must be carried over each of the surfaces S' closing the various leads, and the summation over n will include terms for each of the propagated modes in each of the leads. This same situation will carry through to (III.64), in which we can formally use the same expressions we have already derived, but in which we must now understand that the summation over m involves a summation over each mode of each lead. We see then that a resonant cavity acts like a network with as many pairs of terminals as there are propagation modes of the various leads. A cavity with two leads, each propagating only the dominant mode, acts like a four-terminal network, such as we have discussed in Chapter I. Furthermore, the impedance coefficients are as found in (III.64). We note that the denominator in this expression is to be modified as in (III.72) in case we consider the losses in the walls. Our present result proves the existence of linear relations between the various voltages and currents, and hence justifies the whole treatment of Chapter I in its application to problems of resonant cavities. Beyond that, however, we now see how the impedance coefficients vary with frequency, a feature which we omitted from our discussions completely in Chapter I.

In case there are a number of output leads, we can treat the problem, as we have just seen, like a network with an appropriate number of pairs of terminals. If power is being fed in only through one lead, and in only one mode, however, and if all the other leads and modes are terminated with passive impedances, they will contribute merely to the Q and to the displacement of the resonant frequency of the cavity, and the problem may be handled as that of a cavity with one output. Following back over the argument, we see that we can handle these terms as in the last paragraph. Suppose that a particular mode of the particular lead in which power is being fed in is denoted by the mth, and that each other mode of each lead, say the nth, is terminated by an impedance Z_n. We wish to find the input impedance $V_m/i_m = Z_m$ looking into the mth mode. We have

$$V_n = \sum_a v_{an} \int \mathbf{E}\cdot\mathbf{E}_a dv, \qquad \text{(III.74)}$$

for any value of n, including m. The impedance looking into the nth mode is

$$\frac{V_n}{i_n} = \sum_a \frac{v_{an}}{i_n} \int \mathbf{E}\cdot\mathbf{E}_a dv. \qquad \text{(III.75)}$$

As in (III.64) and (III.65), if we are in the neighborhood of the ath resonant frequency, the term in a in (III.75) will be large and rapidly varying with frequency, while the other terms will lump together to a slowly varying term. Let us assume that we are near the ath resonant frequency, and as in (III.66) lump together all these slowly varying terms, rewriting (III.75) in the form

$$\frac{V_n}{i_n} = \frac{v_{an}}{i_n} \int \mathbf{E}\cdot\mathbf{E}_a dv + Z_{an}, \qquad \text{(III.76)}$$

where Z_{an} is simply defined as the sum of all terms of (III.75) except that with index a. We now have, since the output impedance at the nth mode is Z_n, for $n \neq m$, the relations

$$-Z_n = \frac{v_{an}}{i_n} \int \mathbf{E}\cdot\mathbf{E}_a dv + Z_{an},$$
$$\frac{i_n}{\int \mathbf{E}\cdot\mathbf{E}_a dv} = \frac{-v_{an}}{Z_n + Z_{an}}. \qquad \text{(III.77)}$$

We then wish to find

$$Z_m = \frac{v_{am}}{i_m} \int \mathbf{E}\cdot\mathbf{E}_a dv + Z_{am} \qquad \text{(III.78)}$$

and to get it we must know an accurate value of $\int \mathbf{E}\cdot\mathbf{E}_a dv$. To find this, we use (III.42) again, as in the derivation of (III.55). Replacing the integral $\int (\mathbf{n}\times\mathbf{E})\cdot\mathbf{H}_a da$ by the value found in Section 4, for the case where E is almost exactly proportional to E_a, and replacing the integral $\int (\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da$ by the value (III.62), rewriting the i_n's in that expression from (III.77), we have

$$\left[j\left(\frac{\omega}{\omega_a} - \frac{\omega_a}{\omega}\right) + (1+j)\int \frac{\delta}{2} H_a^2 da + \sum_{n\neq m} \frac{(v_{an}^2/\epsilon_0\omega_a)}{Z_n + Z_{an}}\right] \int \mathbf{E}\cdot\mathbf{E}_a dv + \frac{1}{\epsilon_0\omega_a} \int \mathbf{J}\cdot\mathbf{E}_a dv = \frac{1}{\epsilon_0\omega_a} i_m v_{am}. \qquad \text{(III.79)}$$

Introducing the unloaded Q from (III.59), the modified resonant frequency ω_a' from (III.72), and the external Q from (III.68), we then have

$$\int \mathbf{E}\cdot\mathbf{E}_a dv = \frac{i_m v_{am}/\epsilon_0\omega_a}{j\left(\frac{\omega}{\omega_a'}-\frac{\omega_a'}{\omega}\right)+\frac{1}{Q_a}+\sum_{n\neq m}\frac{1/Q_{\mathrm{ext},\,an}}{Z_n+Z_{an}}+\frac{1}{\epsilon_0\omega_a}\frac{\int \mathbf{J}\cdot\mathbf{E}_a dv}{\int \mathbf{E}\cdot\mathbf{E}_a dv}} \tag{III.80}$$

and, substituting in (III.78),

$$Z_m = \frac{1/Q_{\mathrm{ext},\,am}}{j\left(\frac{\omega}{\omega_a'}-\frac{\omega_a'}{\omega}\right)+\frac{1}{Q_a}+\sum_{n\neq m}\frac{1/Q_{\mathrm{ext},\,an}}{Z_n+Z_{an}}+\frac{1}{\epsilon_0\omega_a}\frac{\int \mathbf{J}\cdot\mathbf{E}_a dv}{\int \mathbf{E}\cdot\mathbf{E}_a dv}}+Z_{am}. \tag{III.81}$$

Equation (III.81) exhibits the input impedance in a mode of the output leads as a resonance term, plus a slowly varying term. In the resonance term, we see the effect on the resonance frequency and the Q of all the types of perturbation which we met in (III.55): the losses in the walls, the losses in the windows, and the losses on account of current in the cavity. We see that each mode of each output lead furnishes a correction to the frequency, and a contribution to Q, just like the value given in (III.68) for a single mode of a single lead. We shall discuss the term involving currents in the next section. In Eq. (III.81), we have a formula correct near the ath resonance frequency, when the resonance term becomes large. We may reasonably assume that the slowly varying term Z_{am} may be approximated, as for instance we saw by comparison of (III.65) and (III.66), by a sum over a of terms like the resonance term of (III.81). Such a sum will have the correct behavior near each resonance. Furthermore, far from resonance the first term in a resonance denominator will be large compared to the others, and this first term is just as given in (III.65), in which we showed that the sum of all the resonance terms except the ath was equal to Z_{a1}. In other words, we are justified in assuming that Z_{am} is approximately given by a sum, over all values of a except a itself, of resonance terms like that in (III.81). The only case in which we may assume that this formula is not very accurate is that in which two resonant modes are close together, so that the breadths of the two resonant peaks are great enough so that they overlap. In such a case, more elaborate methods than we have used are necessary to get valid approximations to the input impedance. If all resonances are well separated, however, (III.81) should be accurate enough for ordinary purposes.

6. Currents within the Cavity

We have not so far considered the contribution of currents in the cavity to the losses and frequency displacement. From either (III.55) or (III.81) we see that this contribution is

$$\frac{1}{Q}-2j\frac{\Delta\omega_a}{\omega_a}=\frac{1}{\epsilon_0\omega_a}\frac{\int \mathbf{J}\cdot\mathbf{E}_a dv}{\int \mathbf{E}\cdot\mathbf{E}_a dv}. \tag{III.82}$$

The simplest application of these terms is the case where the cavity is uniformly filled with a material which conducts according to Ohm's law, with a conductivity σ, and which also has a dielectric constant ϵ, rather than the value ϵ_0 characteristic of empty space. From the conductivity, there is a current density $\mathbf{J}=\sigma\mathbf{E}$, and from the polarization $\mathbf{P}=(\epsilon-\epsilon_0)\mathbf{E}$ there is a current density given by the time rate of change of $\mathbf{P}$, or by $j\omega(\epsilon-\epsilon_0)\mathbf{E}$, if there is a sinusoidal field. In that case, from (III.82), we see that we have

$$\frac{1}{Q}=\frac{\sigma}{\epsilon_0\omega_a},\quad \frac{\Delta\omega_a}{\omega_a}=-\frac{1}{2}\left(\frac{\epsilon-\epsilon_0}{\epsilon_0}\right). \tag{III.83}$$

The first of these indicates greater and greater losses as the conductivity becomes larger, and the second gives the first-order correction to the frequency if the cavity is filled with a dielectric; the correct formula for the change of frequency, as we can see by substituting ϵ for ϵ_0 in (III.45), is that the frequency is proportional to $1/\sqrt{\epsilon}$, from which the frequency change of (III.83) follows at once.

As a next more complicated case, we may imagine that there is conducting or dielectric material distributed through part of the cavity, but not everywhere. In that case, the distribution of **E** and **H** might be violently altered, and in this case our assumption that the field is nearly $\mathbf{E}_a$ in the neighborhood of the ath resonant frequency, on which parts of our deduction are based, would not hold. If the effect of the conductor or dielectric material in the field distribution is not great, however, so that we can approximately replace **E** by $\mathbf{E}_a\int\mathbf{E}\cdot\mathbf{E}_a dv$, (III.82) becomes

$$\frac{1}{Q}-2j\frac{\Delta w_a}{\omega_a}=\frac{1}{\epsilon_0\omega_a}\int(\sigma+j\omega(\epsilon-\epsilon_0))E_a^2 dv.$$

That is, the effect of a conductor or dielectric is large at places where $\mathbf{E}_a$ is large, small where it is small.

A different type of problem is that in which the cavity acts like an oscillator, the currents leading to the generation of power. In such a case, in (III.81), the impedance Z_m would correspond to having a load impedance in the mth mode. That is, we could change the sign of the right side of (III.81), and regard Z_m as an external impedance. A simple transformation then leads to the result

$$-\frac{1}{\epsilon_0\omega_a}\frac{\int\mathbf{J}\cdot\mathbf{E}_a dv}{\int\mathbf{E}\cdot\mathbf{E}_a dv}=j\left(\frac{\omega}{\omega_a'}-\frac{\omega_a'}{\omega}\right)+\frac{1}{Q_a}+\sum_n\frac{1/Q_{\mathrm{ext},\,an}}{Z_n+Z_{an}}. \quad \text{(III.84)}$$

In this expression, the summation over n includes all modes of the output leads; they are all on the same basis. This equation furnishes the fundamental basis for all discussions of microwave oscillators. There are important cases in which the left side can be transformed into a form suggesting lumped constants. In various types of oscillators, such as the magnetron and the klystron, the current is largely confined to a region small compared to a wave-length, and the electric field is likewise confined to such a small region. For instance, in the klystron, both electric field and current are in the region between the grids, which acts almost exactly like a lumped capacity. Thus there is a definite voltage V between the grids. Similarly there will be a current I flowing in the direction opposing the field (if the current is generating power). If we treat the condenser as having an area A, distance of separation between the plates d, then we shall have $E=V/d$, $J=-I/A$, if positive current is in the same direction as positive voltage. We shall furthermore have $\mathbf{E}_a$ a constant over the volume of the condenser, where alone **E** and **J** are different from zero. Thus we shall be able to rewrite the left side of (III.84) in the form

$$-\frac{1}{\epsilon_0\omega_a}\frac{\int\mathbf{J}\cdot\mathbf{E}_a dv}{\int\mathbf{E}\cdot\mathbf{E}_a dv}=\frac{1}{(\epsilon_0 A/d)\omega_a}\frac{I}{V}=\frac{g+jb}{C\omega_a}, \quad \text{(III.85)}$$

in which $g+jb$ is written for I/V, the ratio of current to voltage, and $C=\epsilon_0 A/d$ is the capacity of the condenser. This same formula can be justified in another perhaps more general form, by considering the definition of Q in terms of energy loss. Thus suppose the current, instead of acting as a generator, is acting as a load, and is contributing to the energy loss in the cavity. In this case the term on the left of (III.84) should have a real part which is the contribution of these losses to $1/Q$. Let us suppose that the complex voltage is V, the complex current $I=(g+jb)V$. Then the decrease of energy per unit time is $\frac{1}{2}Re\,IV=\frac{1}{2}g|V|^2$, and the stored energy is $\frac{1}{2}C|V|^2$, where C is the capacity, so defined as to give the stored energy properly in terms of the voltage. The contribution to $1/Q$ is then given by (III.51), from which we see at once that it is $g/C\omega_a$, as we should deduce from (III.85). This formula then holds, so long as it is possible to define a voltage, current, and capacity, satisfying the correct energy relations. Using (III.85), then, the final result is

$$\frac{g+jb}{\cdot C\omega_a}=j\left(\frac{\omega}{\omega_a'}-\frac{\omega_a'}{\omega}\right)+\frac{1}{Q_a}+\sum_n\frac{1/Q_{\text{ext},an}}{Z_n+Z_{an}}. \quad \text{(III.86)}$$

We shall use this formula in the next chapter for giving a general discussion of the properties of microwave oscillators.

7. Perturbation of Boundaries

One more use of (III.42) is frequently very valuable: the problem of the change of frequency of a resonent cavity, when its boundaries are perturbed, as by pushing a small part of the wall in or out. Thus let us start with a cavity, entirely enclosed by a perfect conductor, in which we have found functions $\mathbf{E}_a$, $\mathbf{H}_a$. Now let the wall be pushed into the cavity by a small amount, and let us consider the solution of the problem with the perturbed walls. In the small volume between the original wall and the perturbed one, the final functions $\mathbf{E}$ and $\mathbf{H}$ will be zero. Thus there will be a discontinuity of the tangential component of $\mathbf{H}$ at the perturbed wall. This corresponds to a surface current, and hence, in (III.42) and (III.43), we must include an integral $\int(\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da$ over such a surface, even though it is not part of a surface S'. The perturbed field $\mathbf{H}$ will be very nearly equal to the unperturbed $\mathbf{H}_a$ over the surface, so that the term can be approximated as $\int(\mathbf{n}\times\mathbf{H}_a)\cdot\mathbf{E}_a da$, integrated over the perturbed surface. This may be rewritten $\int-\mathbf{n}\cdot(\mathbf{E}_a\times\mathbf{H}_a)da$. But we may transform this by using

$$\begin{aligned}\text{div}\,(\mathbf{E}_a\times\mathbf{H}_a)&=\mathbf{H}_a\cdot\text{curl}\,\mathbf{E}_a-\mathbf{E}_a\cdot\text{curl}\,\mathbf{H}_a\\&=k_a(H_a^2-E_a^2). \quad \text{(III.87)}\end{aligned}$$

We integrate this quantity over the small volume between the original and perturbed boundaries. The integral of the left side, by the divergence theorem, equals the surface integral of the normal component of $\mathbf{E}_a\times\mathbf{H}_a$ over the boundary. Over the original surface, we may replace the quantity $\mathbf{n}\cdot(\mathbf{E}_a\times\mathbf{H}_a)$ by $(\mathbf{n}\times\mathbf{E}_a)\cdot\mathbf{H}_a$, in which $\mathbf{n}\times\mathbf{E}_a$ is zero over the surface, and hence the contribution vanishes. Over the perturbed surface, the contribution is just $\int-\mathbf{n}\cdot(\mathbf{E}_a\times\mathbf{H}_a)da$, the quantity desired above, since $\mathbf{n}$ is the inner normal into the volume enclosed between the two surfaces. Thus we have

$$\int(\mathbf{n}\times\mathbf{H})\cdot\mathbf{E}_a da=k_a\int(H_a^2-E_a^2)dv, \quad \text{(III.88)}$$

where the surface integral is over the perturbed surface, the volume integral over the volume between the surfaces. If the field $\mathbf{H}$, instead of being $\mathbf{H}_a$, is a constant times this, or $\mathbf{H}_a\int\mathbf{H}\cdot\mathbf{H}_a dv$, then we must multiply the right side of (III.88) by $\int\mathbf{H}\cdot\mathbf{H}_a dv$. We may then introduce this integral into (III.43), finding

$$\begin{aligned}-\epsilon_0\mu_0\omega^2+k_a^2&=-k_a^2\int(H_a^2-E_a^2)dv,\\ \omega^2&=\omega_a^2\left(1+\int(H_a^2-E_a^2)dv\right).\end{aligned} \quad \text{(III.89)}$$

This very valuable formula gives the perturbed frequency ω, resulting from a mode of resonant frequency ω_a, in terms of an integral $\int(H_a^2-E_a^2)dv$ over the volume which is removed from the volume of the resonant cavity by the perturbation of the surface. We see that when the surface is pushed in, the frequency increases if the magnetic field is strong at the part of the surface perturbed, and is decreased if the electric field is strong there. As a simple illustration of this formula, let us start with a cylindrical cavity, in the mode in which $\mathbf{E}$ runs axially from one face to the other, and let us consider the charge in frequency when posts are extended down from the two faces toward each other, as in constructing a klystron. The electric field is strong at the place where the posts are introduced; thus the effect is to decrease the frequency. As another illustration, consider the effect of making a hole in the cylindrical face of such a cavity. Here the magnetic field is strong, the electric field zero; pushing in the wall would increase the frequency, whereas pushing it out decreases the frequency. Finally, there can be simple cases where the effects cancel; thus in the same cylindrical cavity, pushing down the whole top face of the cavity decreases the frequency on account of the center part, where the electric field is large, but increases it on account of the outer part, where the magnetic field is large. A calculation of $\int (H_a^2 - E_a^2)dv$ over the cylindrical face would show that these effects just balance, and actually such a change of dimension, with that particular mode of the cavity, does not change the frequency at all. We shall meet other examples of the use of this theorem later on. It must be remembered that it is strictly correct only for an infinitesimal distortion of the surface.

IV. APPLICATIONS OF THE THEORY OF RESONANT CAVITIES

1. The Tuning of Resonant Cavities

As a first and very instructive example of the application of our general theory of resonant cavities, we consider the following problem: a resonant cavity is provided with a wave guide output of some form, and the guide is closed with a movable short circuiting plunger. We ask, how do the frequencies of the various resonant modes vary with plunger position? This is the problem met when we try to tune a cavity by connecting with a tunable wave guide, and it furnishes a simple example of the general problem of two coupled cavities, one of which can be tuned through the resonant frequency of the other. At the same time, as we shall see later, it has a very close relationship to the problem of determining the input impedance of a cavity, as a function of frequency, by measuring the standing wave ratio and position of the standing wave minimum as functions of frequency. We shall assume for the present that there are no losses in the system, and no damping of any sort. We consider the surface S' of the preceding chapter as being a definite cross section of the wave guide output, between the cavity and the plunger. Let the distance from S' out to the plunger be d. Then the impedance looking out across S', assuming that the guide will propagate only in its dominant mode, is $jZ_0 \tan 2\pi d/\lambda_g$, where Z_0 is the characteristic impedance of the guide, λ_g the guide wave-length. This must be the negative of the input impedance seen looking into the cavity across the same plane, which is given by (III.71). Thus we have

$$\tan \frac{2\pi d}{\lambda_g} = \sum_a \frac{1/Q_{\text{ext},a}}{(\omega/\omega_a)-(\omega_a/\omega)} \qquad \text{(IV.1)}$$

or

$$d = \frac{\lambda_g}{2\pi} \tan^{-1} \sum_a \frac{1/Q_{\text{ext},a}}{(\omega/\omega_a)-(\omega_a/\omega)} + \tfrac{1}{2}n\lambda_g, \qquad n = \text{integer}. \qquad \text{(IV.2)}$$

This equation determines the wave-length (ω, and λ_g, which is of course related to it), as a function of d, the plunger position, or more directly d as a function of the wave-length. The curves determined by (IV.2) are similar to those given in Fig. 12. Far from a resonant frequency, the summation over a is small (equal to Z_a, of (III.71)), so that d is approximately $n\lambda_g/2$, shown by a set of straight lines through the origin in the figure. As the frequency goes through a resonant frequency, however, the summation goes to infinity, changes sign, and again becomes small far from the resonance on the other side; the inverse tangent in the process increases by π, so that d increases by $\lambda_g/2$, or the curve crosses from one of the straight lines to the next. Just at resonance, $d = (n+\frac{1}{2})\lambda_g/2$. In the figure, only two resonances are shown; but actually there will be an infinite set, stretching down to shorter and shorter wave-lengths without limit.

A number of observations can be made about this tuning curve, as we can call it (since it shows how the frequency of the cavity is tuned by moving the plunger). In the first place, a position of the plunger can be found by which the resonant frequency has any desired value. A mode which has, for instance, the frequency ω_1 for one position of the plunger, will tune continuously into a mode with frequency ω_2, simply by moving the plunger. The resonant frequencies are, of course, periodic with plunger position, increase of d by a half guide wave-length bringing the whole set of frequencies back to their original values. We observe next that the resonant behaviors are essentially tied up with the intersections of the dotted lines $d = n\lambda_g/2$, and the dotted lines representing the various resonant frequencies. That is, they come when a resonant frequency of the wave guide itself coincides with a resonance of the cavity, so that we have essentially a coupled system. We notice furthermore that a resonance with a small external Q, or a tight coupling, has the effect of pushing the tuning curve far from the intersection of the dotted lines (as with the resonance ω_1 in the figure) while a resonance with large external Q or loose coupling, like ω_2 in the figure, has the effect of letting the tuning curves approach each other very closely at the resonance. In the case of a loose coupling, then, as we move the plunger, we find that over wide ranges of plunger position, the resonant frequency is almost independent of plunger position; we find definitely non-tuning resonances of this type, which are resonances of the cavity, and also other resonances, which tune greatly, following the dotted lines $n\lambda_g/2$, which are the resonances of the output wave guide, and do not affect the cavity at all. Only when these frequencies coincide do we get appreciable tuning of the cavity resonance. Thus this situation is not suitable for an actual tuner for a cavity resonance. With the tight coupling, however, the cavity resonance tunes strongly over a wide range of tuner positions, so that this is the situation actually desired for a tuner for a cavity.

An interesting insight into the nature of the tuning curves is found from the relation (III.89) of the last chapter, in which we studied the change of resonant frequency of a cavity when we push in a section of wall. From that equation, we find that if the field is large near the movable section of wall (which in this case may be taken to be the plunger) a small displacement will make a large frequency change, while if the field is small it will make a small frequency change. We notice that in the vertical part of the tuning curve, a large displacement makes only a very small frequency change; that means that the field at the plunger, and in the wave guide in general, is very small in this case. In other words, that corresponds to a resonance of the cavity, the field being such that only a small amount of it is located in the wave guide output. On the other hand, in the approximately horizontal parts of the curve, a small displacement makes a large frequency change, showing that the field is strong in the wave guide, as we should expect if it is the guide that is resonating, rather than the cavity.

Suppose we take the intersection of the tuning curve with the straight line $d=\lambda_g/4$. That is, we ask for the resonant frequencies when there is a short circuit a quarter-wave down the line from the surface S'. In this case, there must be an

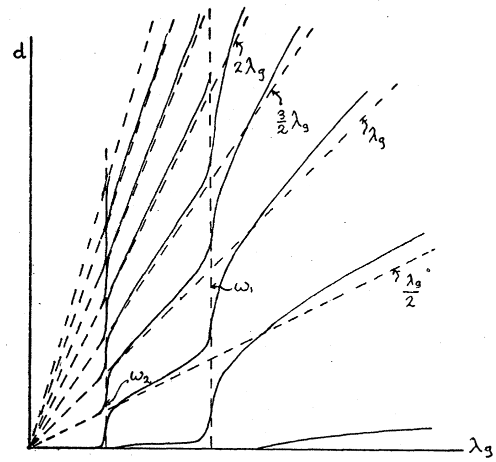

FIG. 12. Tuning curve of a resonant cavity.

open circuit at S'; that is, these resonant frequencies are just the resonances ω_a. Furthermore, from the slope of the tuning curve at these intersections we can find the external Q's. If we differentiate d, in (IV.2), with respect to λ_g, and set the frequency at such a value that $\omega=\omega_a$, so that the ath term of the summation becomes infinite, we easily find

$$\frac{dd}{d\lambda_g}=\frac{(n+1/2)}{2}-\frac{Q_{\text{ext},a}}{\pi}\frac{d\ln\omega}{d\ln\lambda_g}, \qquad \text{(IV.3)}$$

which shows at once, as we have already pointed out, that the vertical part of the tuning curve is very steep for a case of loose coupling, or large external Q, but which also shows that from this slope we can find the external Q directly. Unfortunately, though this gives in principle a way of finding the external Q, it is not accurate in practice, on account of the very large slope, which is hard to measure.

We may now consider a question which has been disregarded until the present: how has the surface S' been chosen? We have stated merely that it is an arbitrary surface in the wave guide output. We could, then, equally well have chosen another surface. If we had done this, however, the distances d to given plunger positions, measured from the new surface S', would have been different. In other words, the tuning curve would have been moved up or down in the figure, by the amount of displacement in S' from its original position. The straight line $d=\lambda_g/4$ would then intersect the tuning curve at different places, so that with the new S' the resonant frequencies would be different; and the slope of the tuning curve at the intersection would be different, so that the external Q would be different. This is the most direct way of seeing the fact, which we mentioned in Chapter III, Section 5, that the external Q's were a function of the position of S' along the line.

It is now natural to ask, is there some particularly correct way of choosing S', which leads to more sensible results than any other? We have drawn the figure in a particular way: with one of the branches of the tuning curve lying close to $d=0$, and others close to $d=n\lambda_g/2$. Clearly a change in S' would change this situation. At least in the neighborhood of one of the resonances, we can always secure this situation, though the same choice of S' will not always be correct for different resonances. Suppose we are considering a resonance of high Q. Near the resonance frequency, by (III.71), the summation can be replaced by its ath term, plus the slowly varying quantity Z_a. If we entirely omitted the resonance term, then the tuning curve would go right through the resonance frequency without showing a resonance phenomenon. If we wished this tuning curve to coincide with the line $d=0$ in the neighborhood of the resonance, we should then have $\tan 2\pi d/\lambda_g = Z_a = 0$. We can do this, on account of the fact that Z_a is a function of the choice of S'. We can give a physical meaning to the process of omitting the resonance term. Suppose that our cavity is tunable, as for instance a wave-meter cavity. Then by tuning it, the resonance frequency ω_a can be shifted around at will. The process of tuning, however, will have relatively small effect on Z_a, made up as this is of contributions from an infinite number of resonant modes. We may then tune the resonance away from the frequency ω in whose neighborhood we are considering Z_a, and the remaining input impedance of the cavity will be Z_a. But now this impedance, as seen across an arbitrary plane, will vary, just as any impedance takes on different values across different planes in the transmission line. Thus if we find that $Z_a=0$ across the plane S', we should then find across a plane distant d from this plane that $Z_a = -j\tan 2\pi d/\lambda_g$. In other words, there is an infinite standing wave ratio (since we have a purely reactive load), and S' is the plane of the standing wave minimum. We may, then, choose S′ to be the plane of the standing wave minimum, when the resonance is tuned away from the frequency at which we are working; then Z_a will be zero, and if we tune the resonance back to the frequency at which we are making our measurements, the input impedance will consist of just one resonance term. We can make this even more correct, if we notice that Z_a will always be a function of frequency, by choosing S' to be a plane whose position depends on frequency, taking at each frequency the position of the standing wave minimum looking into the cavity, when the resonance is tuned out of the way.

2. Measurement of the Properties of a Cavity Resonance

The input impedance looking into a cavity, in which we can no longer neglect the losses, can be written

$$\frac{Z}{Z_0}=\frac{1/Q_{\text{ext},a}}{j[(\omega/\omega_a)-(\omega_a/\omega)]+(1/Q_a)}+\frac{Z_a}{Z_0}, \quad \text{(IV.4)}$$

in the neighborhood of the ath resonance. We shall now ask how the parameters describing this resonance can be found by measurement of the standing wave ratio and position of standing wave minimum looking into the cavity, as a function of frequency. In the first place, if the losses are negligible, so that Q_a is infinite, the standing wave ratio will be infinite at all frequencies, and the position of the standing wave minimum in the line will represent a plane of zero impedance, which could be closed by a short circuiting plunger without change of conditions. In other words, the curve of position of standing wave minimum as a function of guide wavelength is just the same as the tuning curve which we have already shown in Fig. 12. As we go through a cavity resonance frequency, the position of the standing wave minimum rather suddenly shifts by a half wave-length, while between resonances the positions of the standing wave minima move along gradually and regularly.

In case the losses must be considered, the problem is more involved. Let us first consider the form of the curve of impedance *versus* fre-

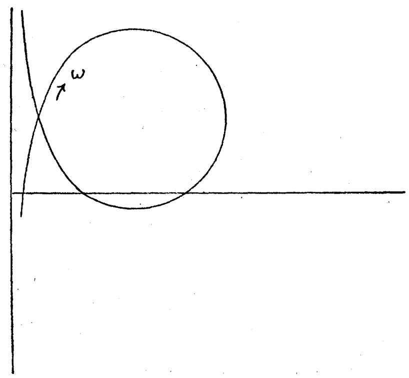

FIG. 13. Impedance of a resonant cavity, for frequencies near resonance.

quency in an impedance plane. If the frequency takes on all values, the first term of the right side of (IV.4) is a circle in the impedance plane; for Z/Z_0 is a bilinear function of the quantity $j(\omega/\omega_a-\omega_a/\omega)$, which takes on only pure imaginary values, and hence traces out a straight line, the imaginary axis, in its own complex plane, so that its transform into the impedance plane must be a circle. Corresponding to frequencies far from resonance, where the denominator becomes very large, the impedance goes to zero, so that the circle passes through the origin; at resonance, the impedance is again real, and we see immediately that $Z/Z_0=Q_a/Q_{\text{ext}}$, at resonance. Thus the two intersections of the circle with the real axis are determined, and since the circle is clearly symmetrical with respect to the real axis, it is completely fixed in position by these two conditions. We see that there is a special case when the loaded Q equals the external Q: the cavity forms a matched load at resonance, so that all power fed down the line is absorbed by the cavity. If the unloaded Q is greater than the external Q, the impedance at resonance will be greater than the characteristic impedance, and if Q_a is less than Q_{ext} the impedance will be less than the characteristic impedance. When now we add the quantity Z_a/Z_0, as a first approximation we simply shift the circle; if Z_a/Z_0 is purely reactive, we shift it vertically, but Z_a will actually have a small resistive component, so that we shift it slightly away from the imaginary axis. On the other hand, Z_a will actually be a function of frequency, though a slowly varying one; as we see from (III.71), for instance, each term of Z_a varies with frequency. Most of the terms come from high resonance frequencies so that the terms of (III.71) can be approximately written in the form $(j/Q_{\text{ext},a})(\omega/\omega_a)$, so that, since each of these varies proportionally to the frequency, the same is the case with Z_a. Thus as the frequency increases, we not only traverse the circle in the impedance plane, in the clockwise direction as we readily verify, but the circle also moves bodily upward. Thus the curve of impedance is similar to that shown in Fig. 13. As the frequency continues to increase, the point representing the impedance will travel upward from the loop shown, and at the next resonance will traverse another loop, and so on indefinitely. We now notice that the procedure of the end of the last section, in which we measured the impedance, not across a fixed plane, but across a plane S' which varied with frequency, such that it always formed a standing wave minimum when the resonance was tuned out of the way, corresponds to disregarding the gradual vertical motion of the circle in the figure above, replacing (IV.4) by the first term, represented by a circle with center on the axis of abscissas, as we first described it.

To find the standing wave ratio and position of standing wave minimum, we wish the plot of the impedance as a function of frequency, not in the impedance plane as in Fig. 13, but in the reflection coefficient plane. This will be as in Fig. 14. From ρ, the magnitude of the reflection coefficient, we can find the standing wave ratio in voltage, power, or decibels, from (I.34), (I.35), or (I.36); it can be calculated by using the relation $r=(Z-Z_0)/(Z+Z_0)$, in combination with (IV.4). We see that the standing wave ratio

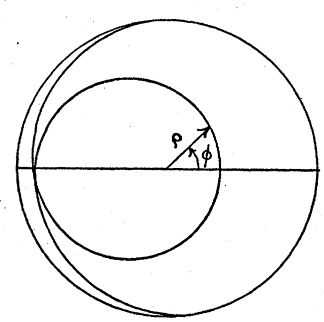

FIG. 14. Impedance of resonant cavity, in reflection coefficient plane.

goes from a very large value off resonance, down to a minimum value at resonance. At the same time the distance d from the plane S' (which is fixed in the form of diagram shown above) out to the position of the standing wave minimum is given by

$$4\pi d/\lambda_g=\phi, \qquad \text{(IV.5)}$$

where ϕ is the angle in the figure. Thus we see that, as we go through the resonance, ϕ increases by 2π, so that d increases by $\lambda_g/2$. In other words, d, the position of the standing wave minimum, as a function of λ_g, behaves in a manner similar to that of Fig. 12, which represents the limiting case of no losses. The situation is quite different, however, in case Q_a/Q_{ext} is less than unity. For in that case, the loop does not encircle the origin in the reflection coefficient plane, so that ϕ, instead of increasing by 2π as we go through a resonance, merely goes through a maximum and minimum, but ends up at almost the same value at which it started. In such a case, the behavior of d, the distance from S' to the standing wave minima, behaves as in Fig. 15.

For the accurate determination of the constants of the resonance, measurement of d, the position of the standing wave minimum, is not nearly as good as a measurement of the magnitude of the standing wave ratio, as a function of frequency. A formulation of the equation of this curve, which has been found convenient in practice, is the following. First, we tune the resonance we are interested in away from the frequency range we are considering, and measure the standing wave ratio as a function of frequency. We choose the plane S' as the plane of the standing wave minimum under these circumstances, so that S' varies with frequency. The impedance as observed across S' is then purely resistive. Remembering, as in (I.37), that the standing wave ratio in voltage is the reciprocal of the value of Z/Z_0 at standing wave minimum, we may define Z_a/Z_0 as σ_1, the reciprocal of the standing wave ratio in voltage off resonance. Next, we tune the resonance back to the point we are interested in, and measure the standing wave ratio on resonance. From (IV.4), this is $Q_a/Q_{\text{ext}}+\sigma_1$. We define this standing wave ratio in voltage as σ_0, so that we have

$$\sigma_0=(Q_a/Q_{\text{ext}})+\sigma_1. \qquad \text{(IV.6)}$$

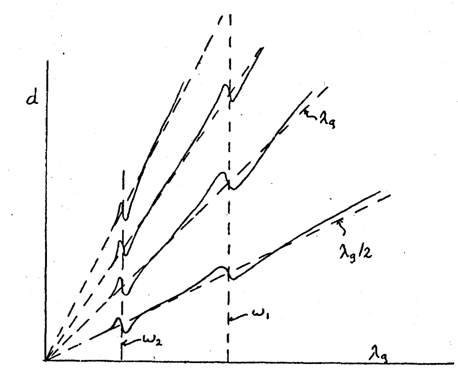

FIG. 15. Position of standing wave minimum as function of wave-length, for resonant cavity with large losses.

There is one qualification to be made to this definition. If we have the case shown in Fig. 14, where Q_a/Q_{ext} is greater than unity (we may call this Case I), then σ_0 will be greater than unity in (IV.6), and it will be the standing wave ratio. On the other hand, if we have the other case, where Q_a/Q_{ext} is less than unity (which we may call Case II), σ_0 will be less than unity, and will be the reciprocal of the standing wave ratio in voltage at resonance. We may distinguish experimentally between the two cases by examining the position of the standing wave minimum d as a function of frequency; if it resembles Fig. 12, we have Case I, while if it resembles Fig. 15, it is Case II.

With the definitions of σ_0 and σ_1, we can then rewrite (IV.4) in the form

$$\frac{Z}{Z_0}=\frac{1}{j\delta+[1/(\sigma_0-\sigma_1)]}+\sigma_1, \qquad \text{(IV.7)}$$

where

$$\delta=Q_{\text{ext}}\left(\frac{\omega}{\omega_a}-\frac{\omega_a}{\omega}\right)\sim 2Q_{\text{ext}}\frac{(\omega-\omega_0)}{\omega_0}. \qquad \text{(IV.8)}$$

Defining the reflection coefficient from (IV.7), we may solve for the magnitude of the reflection coefficient, and find without difficulty

$$|r|=\left[\frac{(\sigma_1-1)^2(\sigma_0-\sigma_1)^2\delta^2+(\sigma_0-1)^2}{(\sigma_1+1)^2(\sigma_0-\sigma_1)^2\delta^2+(\sigma_0+1)^2}\right]^{\frac{1}{2}}. \qquad \text{(IV.9)}$$

From this quantity, we find the standing wave ratio in voltage by the equation

$$SWR(\text{volt})=\frac{1+|r|}{1-|r|} \qquad \text{(IV.10)}$$

and the standing wave ratio in decibels from it by (I.36). The most accurate way to use this equation, in practice, is to observe a complete curve of standing wave ratio as a function of frequency; this curve will resemble Fig. 16. From the frequency of the minimum we find ω_a; from the standing wave ratio at the minimum, we find σ_0. The asymptotic value which σ(db) approaches at some distance from the resonance, or the value which it has at the resonant frequency when the resonance itself is tuned out of the way, gives σ_1. From these values, and (IV.9), we may plot a curve of SWR(db) as a function of δ. Then we choose a horizontal scale, determining frequency in terms of δ, and hence from (IV.8) determining Q_{ext}, which leads to the best agreement with experiment. For rapid testing, we can observe the breadth of the resonance curve at a definite height above the minimum; one such measurement will give Q_{ext}. Various types of charts have been set up, giving convenient points at which to determine this breadth. One convenient chart gives the height of SWR(db) at which the breadth should be determined, in order that the frequency width will give Q_{ext} directly, by the relation $\Delta\omega/\omega = 1/Q_{\text{ext}}$.

By the type of analysis of resonance curves which we have indicated in this section, we can find values of the three parameters Q_{ext}, Q_a, and ω_a, characterizing each resonance of a cavity. It has proved in practice to be a very powerful method of analyzing resonant oscillations, first to use our general theory to indicate the form of the input impedance, in terms of these constants for each resonance; and then to use experimental measurement of input impedance to determine the constants experimentally for the resonances which are actually of interest. Of course, we have given in the preceding chapter general directions from which the values of these constants can be computed purely theoretically, by suitable solutions of Maxwell's equations, and integrations over these solutions. Such calculations have been made in a very few cases, with agreement with experiments. The semi-experimental approach suggested here is very valuable in the much greater variety of problems in which exact calculation is too difficult to attempt.

3. Power Flow Through a Cavity

Another way to investigate the resonant properties of a cavity is to use it as a transmission device, allowing power to flow in through one lead, out through another, and measuring the transmission as a function of frequency, as the frequency is varied through resonance. We shall then show that the breadth of the resonant peak is a measure of the loaded Q of the cavity. To carry out this process, we must feed power into the cavity from an oscillator through a line with much padding or attenuation, as described in Section 15, Chapter I; for only in this case will the output of the oscillator be independent of the load. We note that in this case the input line from the oscillator to the cavity will be practically matched, so that, as we see from Section 5, Chapter III, the contribution of the input lead to the loaded Q will be $1/Q_{\text{ext}}$. As we see from that section, the contribution of the output lead to the loaded Q will be a similar term computed for that lead, if the output is matched as well, which we shall assume that it is. The input impedance into the cavity can be written in the form (IV.7), but we shall neglect the correction term σ_1, which is generally small, and we shall lump the loading by the output lead in with the unloaded Q in the term $1/\sigma_0$. Thus we shall assume that the input impedance is $1/(j\delta+1/\sigma_0)$. We now find, from (I.59), that the power flowing into the cavity will be given by

$$\frac{P}{\frac{1}{2}Z_0 A\bar{A}} = \frac{4R/Z_0}{(R/Z_0+1)^2+(X/Z_0)^2}, \tag{IV.11}$$

where R, X, are the resistive and reactive components of input impedance. From the assumption above, we have

$$\frac{R+jX}{Z_0} = \frac{1}{j\delta+1/\sigma_0}. \tag{IV.12}$$

Substituting from (IV.12) in (IV.11), and carrying out a little algebraic manipulation, we find straightforwardly that

$$\frac{P}{\frac{1}{2}Z_0 A\bar{A}} = \frac{4/Q_a Q_{\text{ext},1}}{\left(\frac{\omega}{\omega_a}-\frac{\omega_a}{\omega}\right)^2+\left(\frac{1}{Q_a}+\frac{1}{Q_{\text{ext},1}}+\frac{1}{Q_{\text{ext},2}}\right)^2}, \tag{IV.13}$$

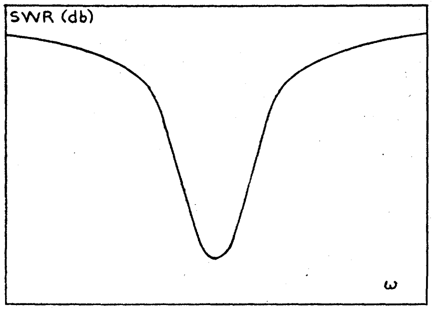

FIG. 16. Standing wave ratio as function of frequency, through resonance.

where $Q_{\text{ext},1}$ is the external Q of the input lead, $Q_{\text{ext},2}$ of the output lead. This equation shows that the power flowing into the cavity follows a resonance curve, the power flow at resonance being

$$\frac{P}{\frac{1}{2}Z_0 A\bar{A}}=\frac{4Q_L^2}{Q_a Q_{\text{ext},1}},$$

where

$$\frac{1}{Q_L}=\frac{1}{Q_a}+\frac{1}{Q_{\text{ext},1}}+\frac{1}{Q_{\text{ext},2}}. \qquad \text{(IV.14)}$$

The maximum value which this factor can have is of course unity, when the cavity is matched to the line; this is the case when $Q_{\text{ext},2}$ is infinite, and when $Q_{\text{ext},1}=Q_a=2Q_L$. We can find the width of the resonance curve at once; the half-power point is found when

$$\frac{\omega}{\omega_a}-\frac{\omega_a}{\omega}\sim\frac{2(\omega-\omega_a)}{\omega_a}=\pm\frac{1}{Q_L}. \qquad \text{(IV.15)}$$

That is to say, the frequency difference between the two half-power points is given by $\Delta\omega/\omega=1/Q_L$. This calculation of power represents that flowing into the cavity. On the other hand, this power will be divided up between the losses in the cavity, and the power flowing out the output lead, in proportion to their contributions to the loaded Q. That is, a fraction

$$\frac{1/Q_{\text{ext},2}}{(1/Q_a)+(1/Q_{\text{ext},2})} \qquad \text{(IV.16)}$$

of the input power will flow out through the output lead. Thus a curve of transmitted power as a function of frequency will have the same form as (IV.13), so that a determination of its width will lead to the loaded Q of the cavity. If the degree of coupling to both leads can be adjusted, as it can if the coupling is through irises of variable sizes, or through a coupling loop whose intrusion into the cavity can be varied, then we may reduce the coupling as far as possible, so that the contribution of the loading of the leads to the loaded Q is negligible; in this limit, the resonance curve will approach a limiting breadth, determined now by the unloaded Q, which can thus be measured by transmission.

4. Properties of a Self-Excited Oscillator

One of the most important uses of the theory of resonant cavities which we have developed is in discussing the properties of self-excited microwave oscillators, such as klystrons and magnetrons. These all operate on essentially the same principles, which we can discuss in general terms without detailed study of the electronic flow within the cavities, which of course leads to their operation. We shall give the general discussion of their circuit properties here, postponing until later a treatment of the electronic motions underlying them. In very simple language, a self-excited oscillator consists of an electronic discharge and a resonant cavity, which plays the part of the "tank circuit" of an ordinary oscillating circuit. The electronic discharge consists of certain currents flowing within the cavity. These currents, by the principles of Chapter III, set up voltages. The voltages in turn are what maintain the discharge. The relation between current and voltage is set by the properties of the discharge, which is always nonlinear, so that for only one amplitude will there be a given ratio between current and voltage. On the other hand, this ratio between current and voltage must also be set up by the resonant circuit. On account of its resonant properties, the correct ratio can be set up only at one frequency, near resonance, so that the resonant cavity has the effect of stabilizing the frequency of oscillation at a definite value.

The theory of the oscillator is essentially contained in Eqs. (III.84) or (III.86). We rewrite (III.86) with the following changes. First, we assume that there is only one mode of the output in which power can escape from the cavity; secondly, we assume that the surface in the output line across which the impedance is measured is such that Z_a is zero, as we have explained in the earlier part of this chapter. Finally, we write the load in terms of its admittance $G+jB$, which is equal by definition to $1/Z$. Then we have

$$\frac{g+jb}{C\omega_a}=j\left(\frac{\omega}{\omega_a'}-\frac{\omega_a'}{\omega}\right)+\frac{1}{Q_a}+\frac{G+jB}{Q_{\text{ext}}}, \quad \text{(IV.17)}$$

which may be written in two parts, the real and imaginary parts of the equation:

$$\begin{aligned}\frac{g}{C\omega_a}&=\frac{1}{Q_a}+\frac{G}{Q_{\text{ext}}},\\ \frac{\omega}{\omega_a'}-\frac{\omega_a'}{\omega}\sim\frac{2(\omega-\omega_a')}{\omega_a'}&=\frac{b}{C\omega_a}-\frac{B}{Q_{\text{ext}}}.\end{aligned} \quad \text{(IV.18)}$$

In this equation, we are assuming a steady oscillation, and a real ω; the case of an exponentially increasing or decreasing amplitude, with a complex ω, will be taken up later, in connection with the problem of starting of oscillations. In (IV.18), we have the relation between the electronic admittance $g+jb$ and the load admittance $G+jB$ and the frequency which is demanded by the circuit properties. In addition to this, however, we have to know certain information about the electronic behavior.

Looking back to Chapter III, the relations (IV.18) give a calculation of the voltages arising from the currents present in the electronic discharge. On the other hand, the purely electronic part of the problem is that which predicts the currents arising from certain voltages; we find this, as we shall show in later chapters, by application of the mechanical laws to the motions of electrons in the assumed fields. If there is a well-defined voltage in the region where the electrons are flowing, as is assumed in the derivation of (IV.18), then we may regard the amplitude of this voltage as an independent variable in solving the electronic problem; and we shall find the electronic current as a function of the voltage. We must note in this that it is the component of current having the same frequency of oscillation as the applied voltage in which we are interested; there will in general be other Fourier components of current, with all multiples of this fundamental frequency, including a constant component, and for the present we can disregard these other Fourier components. The result of the electronic discussion is then a curve giving the electronic current as a function of voltage. Since the discharge is not in general linear, this curve will not in general be a straight line through the origin; and since there is in general a phase difference between voltage and current, there will be both a real and an imaginary component of current, assuming the voltage amplitude to be real. The quantity g is then the ratio of the component in phase with the voltage, to the voltage; b is the ratio of the component out of phase, to the voltage. We may then find as a result of the electronic computations curves for g and b as functions of voltage. No restriction on the shape of these curves is imposed by general considerations; if the problem is linear for very small voltages, g and b will approach constants as the voltage approaches zero, but even this situation does not always hold, notably in magnetrons. We shall find that in important cases g decreases with increasing voltage, the current increasing less rapidly than the voltage. We may well have a situation like that shown in Fig. 17.

Given these curves for g and b as functions of V, which we may arrive at as a result of electronic theory, we may then combine them with (IV.18). The first of those may be rewritten $g/C\omega_a=1/Q_L$, where the loaded Q includes the loading resulting from the conductance G in the output lead. We now see that if G is determined, and hence Q_L, the value of g is fixed, and hence of the voltage V. For the type of curve of g *vs.* V which is shown in Fig. 17, we see that an increase of loaded Q increases the voltage at which the oscillator operates. If the curve of g *vs.* V crosses the axis for a certain finite voltage, as it often does, a sufficiently great decrease of loading will tend to approach this condition. We can never reach it, however, as we see from (IV.18); for with a given resonant cavity, we may decrease G by changing the loading, but we can never get rid of the term $1/Q_a$, so that we can never make g approach zero. In the other direction, by in-

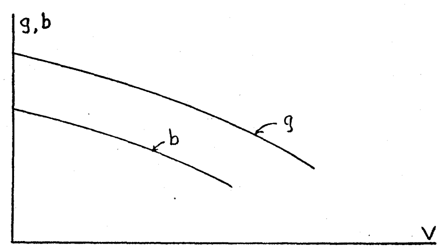

FIG. 17. Electronic conductance and susceptance as functions of voltage, for self-excited oscillator.

creasing the tightness of coupling, which amounts to increasing G or decreasing Q_{ext} so as to decrease the loaded Q, g increases, and if the situation is as in Fig. 17, a quite finite value of g corresponds to zero voltage. Thus a given quite definite tightness of coupling suffices to load the cavity down so that it ceases to oscillate, and at any higher loading than that it cannot operate. For a reflex klystron, this happens at quite ordinary loading; for a magnetron, on the other hand, the curve of g *vs.* V rises much higher, and ordinarily this limit is not reached.

From the conductance G, then, we can fix g, and hence the voltage V. In turn, Fig. 17 shows that b is determined in terms of V. The second equation of (IV.18) then suffices to define the frequency at which the oscillator must operate, for a given load susceptance B. With a given value of voltage, and hence of b, we note that (IV.18) states that the frequency is displaced by an amount proportional to B. This phenomenon is the so-called frequency pulling, resulting from load susceptance. We note that its magnitude is inversely proportional to the external Q; thus the method of decreasing frequency pulling is to decrease the coupling, or increase Q_{ext}. We shall see presently, however, that this has compensating disadvantages. In some cases the quantity b can be controlled by certain subsidiary conditions of the discharge. Thus in a reflex klystron it depends on the reflector voltage, and in a magnetron it depends on the anode current. Changing such a parameter will change the frequency, with a constant load, as we see from (IV.18). In the case of a reflex klystron, this is the phenomenon of electronic tuning and in a magnetron it is the frequency pushing. These phenomena can be used to modulate the frequency of an oscillator electrically, and hence furnish the foundation for a method of frequency modulation.

We have seen how from the external load we can find the voltage V, and the conductance g, of the electronic discharge. We know, however, that the power produced by the electrons, which we may call P_{el}, is equal to $\frac{1}{2}Rei\bar{V}$, where i is the current, V the voltage; or is $\frac{1}{2}gV^2$. Multiplying g as given in the curve of Fig. 17 by V^2, this gives a curve starting off as a parabola for small V, but falling again to zero as g approaches zero. The curve will not have a simple analytic form in general, unless we can derive a simple theory (as we shall find that we can for the reflex klystron) for g as a function of V. We see, however, that as G is changed, varying V, the power will change: as G is increased from a small to a large value, the power will rise from a small value (on account of the low g value at high voltage) through a maximum, and down to a small value again (as the voltage becomes small for high G). It is to be noted that the position of maximum power output cannot be found from any simple rules, such as are used for instance with constant voltage generators; a generator in general does not satisfy the simple postulates of a constant voltage generator. It is not hard to see how a constant voltage generator would behave, however. The voltage across the terminals of such a generator equals its e.m.f. minus the iR drop in the generator. That is, we have $V=E-iR$, so that

$$\begin{aligned} i&=(E-V)/R, \\ g&=i/V=(E/R)(1/V)-(1/R). \end{aligned} \qquad \text{(IV.19)}$$

This is a hyperbola, rising to infinite values when $V=0$ (corresponding to the fact that i is a finite value, E/R, when $V=0$), and crossing the axis when $V=E$. It forms, in fact, not a bad approximation to the curve for the magnetron, where i has large values near $V=0$, and where g rises to very high values in this neighborhood; it is a poor approximation for the reflex klystron, however. For this simple case, the power is given by

$$P_{el}=\tfrac{1}{2}(1/R)(EV-V^2), \qquad \text{(IV.20)}$$

a parabola with maximum at $V=E/2$, or half the voltage at which the value of g goes to zero.

The power produced by the electrons is not all delivered to the external load, on account of the losses in the cavity. We can discuss the losses by the general method given in Section 16 of Chapter I, finding the circuit efficiency, representing the fraction of the electronic power delivered to the load. In terms of it, the output power is then

$$P=\eta_c P_{el}. \qquad \text{(IV.21)}$$

On account of our simple circuit, however, we can give an elementary derivation of the circuit efficiency, at least in case we can neglect the quantity Z_a. The power P_{el} must all be dissipated in the various losses in the circuit, by conservation of energy. These losses are of two sorts, those arising from resistance in the cavity, and from the external load. The power delivered to each of these losses is proportional to its contribution to $1/Q_L$, by the fundamental definition of Q. Since these contributions are $1/Q_a$ and G/Q_{ext}, respectively, the circuit efficiency is given by

$$\eta_c=\frac{G/Q_{\text{ext}}}{1/Q_a+G/Q_{\text{ext}}}=\frac{G}{G+Q_{\text{ext}}/Q_a}. \qquad \text{(IV.22)}$$

This quantity rises from a value of zero when $G=0$ (corresponding to the maximum voltage at which the oscillator can be made to operate) to unity when $G=\infty$, a value which can ordinarily not be reached; at the value corresponding to $V=0$, the circuit efficiency will still be less than unity. Clearly the circuit efficiency will become greater, as Q_a becomes greater (that is, as the losses in the cavity are diminished) or as Q_{ext} becomes smaller (that is, as the coupling between generator and load increases). This bears out our statement that increasing Q_{ext} to decrease the frequency pulling of a generator has a compensating disadvantage: it decreases the circuit efficiency, and hence the output power.

5. Output of Oscillator as a Function of Load

We have seen in the preceding section how the behavior of an oscillator depends on the impedance of the load. The value of G, the load conductance, determines g, and in turn the voltage, the electronic power, and the circuit efficiency, and hence the output power. As G goes from zero to infinity, the power goes from zero (on account of the vanishing of circuit efficiency), through a maximum, and down to zero again (on account of the vanishing of voltage), this second zero value being reached generally at a finite rather than infinite value of G. For any value of G, the frequency in turn is determined by B, as we saw in (IV.18). The value of G fixes b, and this determines the frequency when $B=0$, or for a purely resistive load; changes in B from this value then displace the frequency, by an amount inversely proportional to Q_{ext}. We may then draw curves of constant power, and of constant frequency, in an admittance plane, in which G and B are plotted as abscissa and ordinate. The constant power contours are vertical lines. The constant frequency contours are all obtained from the contour corresponding to $\omega=\omega_a'$ by vertical displacement. That contour is determined by the simultaneous solution of the two equations $g/C\omega_a=1/Q_a+G/Q_{\text{ext}}$, $b/C\omega_a-B/Q_{\text{ext}}=0$. Each value of G leads to a value of g; by Fig. 17, this leads to V, hence to b, and hence to B. We may write the equations

$$G=Q_{\text{ext}}(g/C\omega_a-1/Q_a), \quad B=Q_{\text{ext}}(b/C\omega_a). \qquad \text{(IV.23)}$$

Thus the curve of B *vs.* G is like that of b *vs.* g, except for a change of scale given by the factor $Q_{\text{ext}}/C\omega_a$, and a horizontal displacement given by the term $-Q_{\text{ext}}/Q_a$. Often in practical cases a good enough approximation to this relation is a straight line, which is not in general horizontal. The frequency contours are then such a family of parallel straight lines.

A plot of contours of constant power and constant frequency in the reflection coefficient plane is generally called a Rieke diagram. Since the transformation from admittance plane to reflection coefficient plane is a bilinear transformation, the contours will be circles. The power contours all pass through the point of infinite admittance, or zero impedance, and are all parallel to the line of zero resistance at that point; thus in the reflection coefficient plane they will be circles tangent to unit circle at the point of zero impedance. The frequency contours also pass through this point, but at a different angle. Thus the Rieke diagram has the appearance shown in Fig. 18. It should be remembered that this is on the assumption that $Z_a=0$; that is, that we are measuring load admittance across a

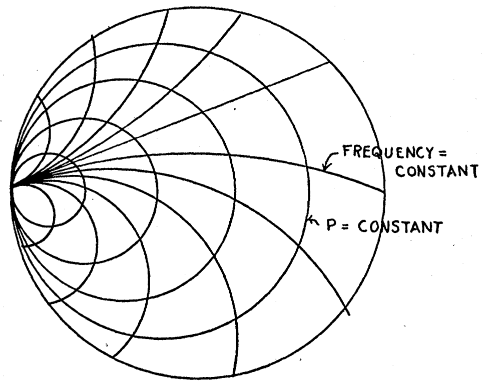

FIG. 18. Idealized Rieke diagram.

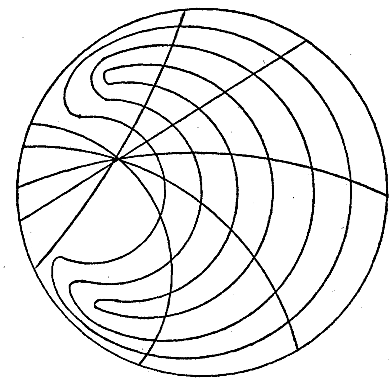

FIG. 19. Actual Rieke diagram.

plane in the output line across which the standing wave minimum is located when the resonance is removed, as by tuning out of the way; this plane, we remember, is a function of frequency. It is a plane which is electrically a whole number of half wave-lengths away from the electronic discharge, so to speak. If we take a Rieke diagram across a fixed plane at a distance from the tube, there are various complications which we shall not go into; they result in having the frequency contours cross in the neighborhood of a point inside the unit circle, as shown in Fig. 19. We must also remember that we cannot in general choose a plane across which the real as well as the imaginary part of Z_a vanishes; there are inevitably losses included in this term. When we analyze the effect of these losses, we find that they affect the dependence of both electronic power and circuit efficiency on G and B. The result is that the output power no longer depends on G alone, but also on B, in such a way that the power contours are deformed in the way shown in Fig. 19.

The Rieke diagram, or corresponding plot in the admittance plane, determines the power and frequency of operation of an oscillator when operating into a given load. Often, however, the load may have an admittance which is not constant, but which is a function of the frequency. In such a case, we cannot specify the load admittance in advance. To find the operating point, we may draw the locus of the load admittances for different frequencies in the admittance plane, draw the contours of constant frequency as before, and ask for what frequency the two coincide. If the load is a resonant cavity with a resonant frequency near that of the oscillator, its input impedance will be like Fig. 13, and the corresponding reflection coefficient like that of Fig. 14. The corresponding admittance will be like that shown in Fig. 20. We see that there can well be more than one point of coincidence of frequencies on the two curves. In other words, it can well happen that the oscillator has a choice of operating in two different frequencies, with a single load. This is closely related to the existence of two resonant frequencies when two circuits of about the same resonant frequency, the tank circuit of the oscillator and the external load, are coupled together. The weaker the coupling, or the larger the external Q, the smaller will be the frequency separation of the two modes of oscillation.

In Fig. 20, the straight lines represent the curves of B *vs.* G for constant frequency, as described in the preceding paragraphs. The frequencies are numbered from 1 to 10. The curve represents an admittance curve for the external resonant load, as a function of frequency, again with the frequencies numbered. It is clear that the two sets of frequencies coincide, so that a resonance exists, for the frequencies labeled 4, 5, and 6. It can be shown that of these resonances, one is unstable, and would not really exist.

6. Starting of an Oscillator

We have spoken of the stable operation of an oscillator, in which the frequency is real. It is of interest to ask, however, how the oscillations build up from zero amplitude. For a short time

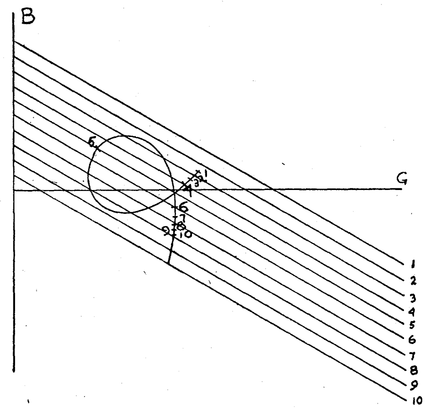

FIG. 20. Admittance plane, for operation of self-excited oscillator into resonant load.

interval during the build-up, we may assume that the amplitude is increasing exponentially with the time, so that formally we may treat the frequency as being complex, the imaginary term representing the exponential increase, and thus having the opposite sign to that which we have previously used in discussing exponential damping, and the dissipation of power in a cavity. Thus suppose $\omega=\omega_1-j\omega_2$, where ω_1 is the real frequency, and where the amplitude is building up according to the exponential $e^{\omega_2 t}$. Substituting this value in (IV.17), we see that

$$2\frac{\omega_2}{\omega_a}=\frac{g}{C\omega_a}-\frac{1}{Q_L}. \quad \text{(IV.24)}$$

We can see the interpretation of this equation if we use a diagram like that of Fig. 17. Plotting $g/C\omega_a$ against V, we see that the difference between this curve and the straight line at height $1/Q_L$ determines the rate of increase of the voltage. Starting at a given voltage, in the case shown in Fig. 21, where the curve of $g/C\omega_a$ lies above that of $1/Q_L$, the voltage will then increase, more or less exponentially, with the time. As the curves then draw closer together (assuming that the curve of $g/C\omega_a$ slopes downward), the rate of increase will flatten off, until finally a steady state will be reached when the two curves intersect. Clearly if the curve of $g/C\omega_a$ slopes upward, the amplitude will increase without limit, and no stable operation will be possible. Also clearly if the curve of $1/Q_L$ lies entirely above that of $g/C\omega_a$, ω_2 will be negative instead of positive, so that even if we start with a finite voltage amplitude, the amplitude will decrease exponentially, and the electronic discharge will act like a load, not a generator.

Clearly the voltage as a function of time is not a simple exponential, but we can give an analytic evaluation of the relationship. We note that if instantaneously the voltage is increasing as $e^{\omega_2 t}$, we have $\omega_2=d\ln V/dt$. Substituting from (IV.24), and integrating, we have

$$t-t_0=\int_{V_0}^{V}\frac{dV}{\frac{\omega_a V}{2}\left(\frac{g(V)}{C\omega_a}-\frac{1}{Q_L}\right)}, \quad \text{(IV.25)}$$

where we assume that $V=V_0$ when $t=t_0$. Over a range of voltage for which g may be treated as a constant, this shows that t varies logarithmically with voltage, or voltage exponentially with time, as we have already seen. This has a bearing on the initial process of build-up. With a curve like that of Fig. 21, in which g approaches a finite value when $V=0$, we have this situation of exponential build-up for small voltages. In other words, it would take an infinite time to build up from zero voltage; all we can do in a finite time is to start with an already existing voltage, and amplify that. In practice, in oscillators having such a characteristic, like the reflex klystron, the oscillation starts from the noise, or fluctuation, voltage always present in an electronic discharge. It is interesting to note that the rate of build-up is greater, the greater the loaded Q, or the greater the final voltage. Quite a different case is that of Eq. (IV.19), in which the current is a linear function of the voltage, so that the conductance becomes infinite at zero voltage. In that case, the term in $1/V$ in $g(V)$ cancels the factor V in the denominator of (IV.25), so that for small values of V the integrand is constant, and V increases linearly with time. In this case it is not necessary to start from noise. The characteristic of having a finite current for zero voltage is obviously discontinuous; for it is also naturally possible to have zero current for zero voltage. Such a situation is found approximately in the magnetron, where the state of zero current is inherently unstable, and breaks at the slightest provocation into a state of large current. This then builds up linearly, finally approaching asymptotically to its limiting value. In fact, if we insert the value (IV.19) for g, and integrate (IV.25), setting $V=0$ when $t=0$, we find at once that

$$V=\frac{E}{RC\omega_a}\frac{1}{(1/RC\omega_a+1/Q_L)}$$
$$\times\left\{1-\exp\left[-\frac{\omega_a}{2}(1/RC\omega_a+1/Q_L)t\right]\right\}. \quad \text{(IV.26)}$$

That is, the voltage increases in a way like that of the voltage in a condenser being charged through a resistance in ordinary circuit theory. The time of build-up is related to the loaded Q, as modified by the term $1/RC\omega_a$, which acts like an additional form of loading as far as the mathematics is concerned. Clearly, then, a magnetron with a high loaded Q, or with a low frequency, will take a relatively long time to reach its full voltage.

7. Experimental Investigation of Electronic Admittance

In Fig. 17 we postulated a form for the curves of g and b as functions of V, and have asked how the oscillator would behave with such values. In the next chapter we shall see that in simple cases, like the reflex klystron, we can actually calculate these curves, and obtain a relatively complete theory of their operation. With more complex cases like the magnetron, however, such calculation is extremely difficult, if not impossible. In these cases, we can work backward from observed operating behavior, and find the curves of g and b from experiment. We ordinarily measure two quantities, power and frequency, as functions of the load admittance $G+jB$. The output power is of course $\eta_c(\frac{1}{2}gV^2)$. Knowing the load admittance and the circuit constants, we can find g from (IV.18), and η_c from (IV.22). Measuring the power we can then compute V, the voltage. Hence we get a curve of g as a function of voltage. Next we measure the lines of constant frequency in the admittance plane. From these, as in Section 5, we can find b as a function of g, and hence as a function of voltage. Thus the complete information can be found experimentally. The curves so found can then be used in (IV.25) to investigate the build-up of the oscillations. Unfortunately comparatively little use has been made of this method for investigating the characteristics of microwave oscillators. Rieke and his collaborators have used it to some extent for magnetrons, but no thorough study has yet been made of the characteristics of any type of microwave oscillator as a function of the various parameters, such as d.c. currents, voltages, etc., which control the oscillation. In the next chapter we shall, however, find the general form of behavior to be expected, for the reflex klystron and the magnetron.

V. ELECTRONICS OF THE REFLEX KLYSTRON AND MAGNETRON

1. The Reflex Klystron

The resonant cavity of the reflex klystron, or reflex oscillator, is shaped fundamentally like a short length of coaxial line, with a gap in the inner conductor (Fig. 22). This gap acts like a lumped capacity, so that in the mode in which it operates, the electric lines of force run between the faces of the gap, and the magnetic lines are in circles surrounding the axis of the cavity. As a result of the loading by the capacity, the length of the cavity is much less than the half wave-length which it would be if the inner conductor were continuous. On the other hand, the diameter of the cavity is also much less than it would be for a cylindrical cavity without inner conductor, operating in the corresponding mode, with **E** along the axis, **H** in circles around it. This may be easily seen by considering how the frequency would change if we started distorting such a cylinder by allowing posts to protrude from the opposite circular faces, and approach each other. The posts would push into the cavity at points of high electric field; thus, by Eq. (III.89), they would have the effect of decreasing the resonant frequency, and to restore it to its original value we should have to decrease the diameter of the

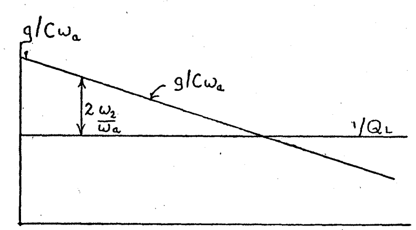

FIG. 21. Conductance curve for starting of oscillator.

cavity, thereby pushing in the walls in a region of high magnetic field. The whole cavity is then rather small compared to a half wave-length, particularly so if the gap between posts is narrow and its capacity is high, so that even more the gap itself is small compared to a half wave-length, and hence the field in its immediate neighborhood can be treated rather accurately by electrostatics. It forms, in other words, a condenser, whose capacity C can be easily found. For a small gap and high capacity, the resonant frequency of the cavity can be rather accurately computed by a simple approximation of assuming a length of coaxial line to be terminated by a capacity.

The two condenser plates, instead of being solid, are made of grids, so that the effect of the oscillation is to impress an r-f voltage between these grids. Thus the resonator forms the simplest form of diode, in the microwave region. An electron gun then shoots a beam of electrons, of a fixed energy, at the pair of grids. These electrons are speeded up or slowed down in passing through the grid system, depending on the phase of the r-f field. The resulting velocity modulation results, after passing through a certain drift distance on the other side of the gap, in a bunching of the electrons: the electrons which have gained energy speed ahead, those which have lost energy lag behind, until at a certain distance they meet and form periodic bunches. In the ordinary klystron, a second resonant cavity and grid system, called the catcher, is located at such a distance from the bunching cavity that the bunching is formed approximately at the grid system of the catcher. If an r-f voltage is impressed on the catcher, the bunches of electrons will either deliver energy, or absorb it, from the cavity oscillations. If they deliver energy, they will maintain the oscillation, and by a proper feedback system between the two cavities the device will act like an oscillator, much as in an ordinary triode oscillator, the buncher fulfilling the function of the grid circuit, the catcher of the plate circuit. On the other hand, if the phase is such that the electrons absorb energy from the catcher, the oscillation will not be maintained, but the electrons will absorb energy on the average in the catcher. By arranging a number of cavities in succession,

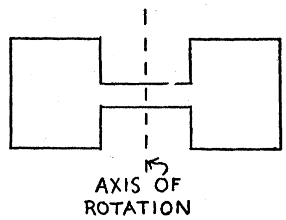

FIG. 22. Schematic diagram of klystron cavity.

with suitable distances, and phase separations between, it is possible to have the bunches of electrons reach each grid system in such phase as to absorb energy, and thereby to gain more and more energy with each cavity which they traverse. This device is the linear accelerator, by which r-f power is converted into d.c. power, in the production of high energy particles. Our present interest, however, is the inverse of this, or the oscillator, in which d.c. power is converted into r-f.

In the reflex oscillator, one cavity and grid system is made to fulfill the function of both buncher and catcher, by having an electrode, called the reflector or repeller, negatively charged, which reverses the electrons after traversing the grids once, and causes them to traverse them again in the opposite direction. The time of transit from the grids back to the grids again depends not only on the dimensions, but also on the repeller or reflector voltage. If this is adjusted properly, the electrons re-enter the grid in such phase as to deliver power, and the oscillations are maintained, the tube operating as a microwave generator. On the other hand, for other adjustments, the electrons enter in such phase as to act as a load, and the oscillation cannot maintain itself; such operation can be observed only by feeding power into the cavity from outside. We find that there are a number of different values of reflector voltage for which oscillation is possible; these differ by such amounts that the transit times of the electrons in the drift space differ by whole periods of the r-f. The power shows a maximum for each of these reflector voltages; they are called different electronic modes of the oscillator. As the voltage is changed from each of these values, the power decreases, and in between them oscillation does not occur. At the same time that the power changes, the phase of the bunched electrons

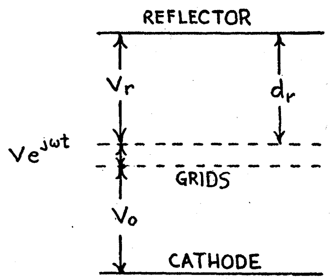

FIG. 23. Schematic diagram of reflex klystron electrodes.

introduces a reactive component of r-f current, which results in a change of frequency of operation of the tube. This is the electronic tuning which makes the reflex oscillator such a convenient power source for any purpose requiring frequency modulation.

We shall now examine the details of this electronic process. Our fundamental object, as is clear from the preceding chapter, must be to determine r-f current, as a function of r-f voltage. The voltage is simply that which is impressed on the grids. To find the r-f current, we shall examine the current carried by the electrons which have been bunched by traveling through the drift space. We shall find that superposed on their d.c. current is an r-f fluctuation. It is this which we must compute, and which forms the r-f current which we must use in computing g and b. We shall find that we can get explicit formulas for those quantities, in a certain limiting case of small amplitude operation, so that we can set up a theory of the operation of the reflex klystron, along the lines of that sketched in the preceding chapter.

2. Electronic Operation of the Reflex Klystron

In Fig. 23 we show the various electrodes of the reflex klystron, in a schematic manner: the cathode, from which the electrons are emitted, being accelerated by a d.c. voltage V_0 before reaching the grid system; the grids, on which the r-f voltage $Ve^{j\omega t}$ is impressed, by the oscillations in the resonant cavity; and the reflector, which repels the electrons, the retarding voltage between grid system and reflector being V_r. We shall treat V_0 and V_r as positive numbers, in case the voltages have the usual sign, and shall take V as positive when it slows down the electrons between the grids. We shall make a number of simplifying assumptions. We assume a one-dimensional problem, neglecting all transverse motions of the electrons, or departures of the fields from the one-dimensional case. We assume the r-f voltage amplitude V to be small compared to the incident voltage V_0 of the electrons, so that we can neglect powers of the ratio V/V_0. We assume the grid separation to be so small that the transit time through the grids can be neglected (it is not hard to remove this restriction, and investigate the effect of transit time). Furthermore, we neglect the possibility of collisions of the electrons with the grids, with their consequent loss from the beam; and we neglect multiple transit electrons, which are not caught by the electrodes after their first transit through the grids.

Let us now consider the motion of an incident electron which strikes the grids at time t_0. If e is the magnitude of the electronic charge ($=1.60\times 10^{-19}$ coulomb), and m is its mass ($=9.0\times 10^{-31}$ kgm), its incident velocity v_0 (in meters per second) is given by

$$\tfrac{1}{2}mv_0^2 = eV_0. \qquad \text{(V.1)}$$

After emerging from the grids it will have a kinetic energy $e(V_0 - V\cos\omega t_0)$, so that its velocity v_1 will be given by

$$\tfrac{1}{2}mv_1^2 = e(V_0 - V\cos\omega t_0). \qquad \text{(V.2)}$$

Solving for v_1, expanding the square root by binomial expansion, and rejecting higher powers of V/V_0 than the first, we have

$$v_1 = v_0[1 - \tfrac{1}{2}(V/V_0)\cos\omega t_0]. \qquad \text{(V.3)}$$

The electron now emerges, with velocity v_1, into the drift space, where it is acted on by a decelerating force of eV_r/d_r. That is, its equation of motion in the drift space is

$$m\frac{dv}{dt} = \frac{-eV_r}{d_r}, \quad v = v_1 - \frac{e}{m}\frac{V_r}{d_r}(t - t_0). \qquad \text{(V.4)}$$

The electron will return to the grids at the time t_1 at which its velocity has reversed, or has become $-v_1$. That is, we have

$$t_1 = t_0 + \frac{2mv_1 d_r}{eV_r} = t_0 + \frac{2mv_0 d_r}{eV_r}\left(1 - \frac{1}{2}\frac{V}{V_0}\cos\omega t_0\right). \qquad \text{(V.5)}$$

We shall define the average transit angle of the electrons, from the grid back to the grid again, as θ. This is ω times the time required for this transit, for the case $V=0$, and is given by

$$\theta = 2m\omega v_0 d_r / eV_r. \tag{V.6}$$

In terms of θ, we may rewrite (V.5) in the form

$$\omega t_1 = \omega t_0 + \theta - \tfrac{1}{2}(\theta V/V_0)\cos\omega t_0. \tag{V.7}$$

We are now ready to find the current as it returns to the grids. Let the cathode current be I_0, which we take to be positive when electrons flow from the cathode, or positive current flows to it. Then the number of electrons per second striking the grids from the cathode is

$$dn/dt_0 = I_0/e. \tag{V.8}$$

These same electrons strike the grid in the reverse direction on their return from the reflector in a time interval dt_1. Thus the returning current is

$$-\frac{dn}{dt_1} = \frac{I_1}{e} = -\frac{dn}{dt_0}\frac{dt_0}{dt_1} = \frac{-I_0}{e}\frac{dt_0}{dt_1}. \tag{V.9}$$

Determining dt_1/dt_0 from (V.7), we see that the returning current is

$$I_1 = -I_0/(1+\tfrac{1}{2}(\theta V/V_0)\sin\omega t_0). \tag{V.10}$$

On account of the sinusoidal function in the denominator of (V.10), we see that I_1 has an r-f component of frequency ω. Equation (V.10) is not, however, a convenient formula from which to determine the r-f component, for it expresses I_1 in terms of t_0 rather than in terms of t_1, the time at which the electrons return. We can easily get around this difficulty as follows. To find the r-f component of I_1, the standard procedure is to multiply by $e^{-j\omega t_1}$, and average over a complete period. That is, we have

$$\begin{aligned} i &= \frac{\omega}{2\pi}\int_{\text{period}} e^{-j\omega t_1} I_1(t_1)dt_1 \\ &= -\frac{\omega}{2\pi}\int_{\text{period}} I_0\frac{dt_0}{dt_1}e^{-j\omega t_1}dt_1 \\ &= -\frac{\omega I_0}{2\pi}\int_{\text{period}} e^{-j\omega t_1}dt_0. \end{aligned} \tag{V.11}$$

In (V.11) we have converted our integral into one over t_0 instead of t_1. Inserting t_1 from (V.7), we have

$$i = -\frac{\omega}{2\pi}I_0\int_{\text{period}} \exp\left[-j(\omega t_0+\theta-\tfrac{1}{2}(\theta V/V_0)\cos\omega t_0)\right]dt_0. \tag{V. 12}$$

We shall introduce the abbreviations

$$z = -\frac{\theta}{2}\frac{V}{V_0}, \quad \omega t_0 = \phi. \tag{V.13}$$

Then (V.12) becomes

$$i = -\frac{I_0}{2\pi}e^{-j\theta}\int_{\text{period}} e^{-j(\phi - z\cos\phi)}d\phi. \tag{V.14}$$

Using Sommerfeld's integral relation*

$$J_n(z) = \frac{j^{-n}}{2\pi}\int_0^{2\pi} e^{jz\cos\phi}e^{jn\phi}d\phi \tag{V.15}$$

and

$$J_{-n}(z) = (-1)^n J_n(z), \tag{V.16}$$

where n is integral, we then have

$$i = I_0 e^{-j(\theta - 3\pi/2)}J_1(z). \tag{V.17}$$

In Eq. (V.17) we have the formula for the r-f current amplitude; the complex current is $ie^{j\omega t}$, associated with the complex voltage $Ve^{j\omega t}$. The sign is so chosen that when the current opposes the voltage, or when the electrons are acting as a generator, i and V are in the same phase. We recall that (V.17) involves a number of approximations, as we have enumerated above.

In addition to the r-f current i of (V.17), we shall want the r-f admittance introduced across the grids by the electrons. This is

$$g + jb = \frac{i}{V} = \frac{I_0}{V_0}\frac{\theta}{2}e^{-j(\theta - 3\pi/2)}\frac{J_1(z)}{z}, \tag{V.18}$$

* See Jahnke-Emde, p. 149.

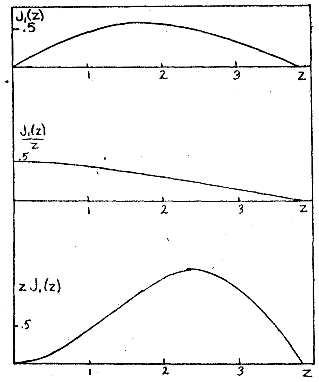

FIG. 24. Dimensionless representation of current, admittance, and power of reflex klystron, as function of voltage.

where we have used (V.13). Separating real and imaginary parts, we have

$$g=\frac{I_0}{V_0}\frac{\theta}{2}\cos\left(\theta-\frac{3\pi}{2}\right)\frac{J_1(z)}{z},$$
$$b=-\frac{I_0}{V_0}\frac{\theta}{2}\sin\left(\theta-\frac{3\pi}{2}\right)\frac{J_1(z)}{z}. \quad \text{(V.19)}$$

We shall also need the electronic power P_{el}, which is $\frac{1}{2}gV^2$, or is

$$P_{el}=I_0V_0\frac{\cos\left(\theta-\frac{3\pi}{2}\right)}{\theta}zJ_1(z). \quad \text{(V.20)}$$

Remembering that I_0V_0 is the input power, we may write the electronic efficiency as the quantity (V.20) with the factors I_0V_0 left out. The three functions $J_1(z)$, $J_1(z)/z$, and $zJ_1(z)$ then represent in a dimensionless way the r-f current, admittance, and power, as functions of the r-f voltage, which is represented dimensionlessly by z. We show these three functions in the curves of Fig. 24, and observe that the middle one, proportional to g, is similar to the schematic curve which we gave in Fig. 17. We notice from (V.19) that the curves of g and b as functions of V are both of the same form; for we have

$$b=-g\tan\left(\theta-\frac{3\pi}{2}\right). \quad \text{(V.21)}$$

That is, the two quantities are proportional to each other, the constant of proportionality being a function of θ, the transit angle. This in turn is a function of the reflector voltage, as we see from (V.6).

3. Power and Frequency of Reflex Klystrons

From the curves of g and b as functions of r-f voltage, which we have just derived, we can carry out a discussion of the power delivered by the tube as a function of load, and of the frequency of operation, as in the preceding chapter. The curve of b *vs.* g, which determines the nature of the frequency contours in the admittance plane, or in the Rieke diagram, is a straight line, as we have just seen in (V.21); furthermore, the slope varies with θ, or with reflector voltage. We shall now show that we can summarize the information about both power and frequency, for all values of θ, in a single simple diagram. This is a figure in which g is plotted as abscissa, b as ordinate. First we can draw a line of constant θ in this space. By (V.21), it is a straight line through the origin, with a slope of $-(\theta-3\pi/2)$. We notice that a horizontal line corresponds to $(\theta-3\pi/2)=2\pi$ times an integer, or to

$$\theta/2\pi=n+\tfrac{3}{4}, \quad \text{(V.22)}$$

where n is an integer. We shall see in a moment that these values of θ correspond to power maxima of the various electronic modes, and we shall refer to the corresponding modes as the $1\frac{3}{4}$, $2\frac{3}{4}$, etc., modes. We observe that as the reflector voltage increases, θ decreases, or the line rotates counterclockwise, or in a positive direction.

Next we can draw contours of constant power. First let us consider the curve corresponding to $z=0$, or vanishing r-f voltage. This curve is obviously a contour of zero electronic power, and may be called the small signal curve. Since it can be shown that $J_1(z)/z$ approaches $\frac{1}{2}$ as z approaches zero, the equation of this curve is

$$g=(I_0\theta/4V_0)\cos(\theta-3\pi/2),$$
$$b=-(I_0\theta/4V_0)\sin(\theta-3\pi/2). \quad \text{(V.23)}$$

That is, if we let $r=(g^2+b^2)^{\frac{1}{2}}$, the length of the radius vector, we have

$$r=(I_0\theta/4V_0) \quad \text{(V.24)}$$

as the equation of the small signal curve, in

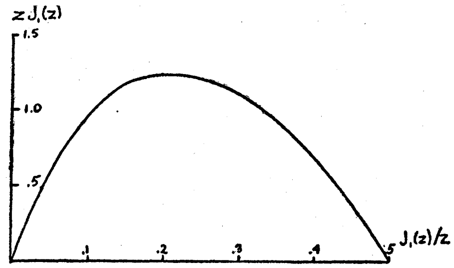

FIG. 25. Dimensionless representation of power as function of admittance, for reflex klystron.

polar coordinates. It is then a spiral, the larger loops corresponding to larger θ, or smaller reflector voltage. We may now consider how the electronic efficiency varies as we go along a radius, from the origin to the small signal spiral. We note that the radius vector is proportional to $J_1(z)/z$, and that the electronic efficiency is proportional to $zJ_1(z)$. To show the relationship between these, we plot $zJ_1(z)$ as a function of $J_1(z)/z$, in Fig. 25. We see that as we go out from the origin to the small signal spiral, the electronic efficiency increases from zero to a maximum, which it reaches at a value equal to about 43 percent of the radius of the spiral, and then decreases to zero again at the spiral. Further, on account of the factor cos $(\theta-3\pi/2)$ in (V.20), the electronic power decreases as θ goes in either direction from the value (V.22) representing a horizontal line; and the maximum power decreases as θ, or n increases. The power is negative for negative g; that is, in the left half plane the tube is a passive load, not an active oscillator.

Before we can draw contours of constant power, we must consider the circuit efficiency. We saw in (IV.22) that this depends on G, or on g. It is zero for $G=0$, or for

$$g=C\omega_a/Q_a. \qquad \text{(V.25)}$$

As G increases, the circuit efficiency increases, becoming unity for large G. The vertical line denoted by (V.25) will then be a contour of zero power, and tube operation can only occur to the right of this line, between the line and the spiral. We can now compute the power at any value of g and b, using (V.19) to find $J_1(z)/z$, from this finding $zJ_1(z)$, and getting the electronic power from (V.20), and the circuit efficiency from (IV.22). In Fig. 26 we show contours of constant efficiency, for the $1\frac{3}{4}$, $2\frac{3}{4}$, and $3\frac{3}{4}$ modes, computed for a case similar to those met in practice. The contours are simple to interpret. As we approach either the vertical line of zero circuit efficiency, or the small signal spiral, the efficiency goes to zero, and it reaches a single maximum in the middle of the operating range, approximately on the horizontal axis. For the modes of small n values, the maximum efficiency is low because the maximum comes for low g, where the circuit efficiency is low; for large n values, the maximum efficiency is low because of the factor θ in the denominator of (V.20). In between, there is an n value for which the peak efficiency has its largest value, in this case the $2\frac{3}{4}$ mode. Clearly the question of which mode has the highest power will depend on the value of $C\omega_a/Q_a$; the smaller it is, the higher the efficiency of all modes, but the greatest improvement will come about in the modes of small n value.

We may now use these figures to discuss the operation of the oscillator. First we consider operation into a matched load, and ask what will be the power and frequency as functions of reflector voltage. Combining Eqs. (IV.18), (V.19), and (V.21), we have

$$g=\frac{I_0}{V_0}\frac{\theta}{2}\cos\left(\theta-\frac{3\pi}{2}\right)\frac{J_1(z)}{z}=C\omega_a\left(\frac{1}{Q_a}+\frac{G}{Q_{\text{ext}}}\right),$$
$$\frac{2(\omega-\omega_a)}{\omega_a}=-\tan\left(\theta-\frac{3\pi}{2}\right)\left(\frac{1}{Q_a}+\frac{G}{Q_{\text{ext}}}\right)-\frac{B}{Q_{\text{ext}}}. \qquad \text{(V.26)}$$

For a matched load, we have $G=1$, $B=0$. Thus by the first equation of (V.26) the operation will be at points of a vertical line, $g=$constant. As the reflector voltage changes, θ changes, and the intersection of the radial line corresponding to the reflector voltage, and the vertical line corresponding to the value $G=1$, will give the operating point. As the reflector voltage changes, the power will go from the maximum value corresponding to the horizontal axis in the figure, down to zero, and the tube will go out of oscillation, starting up again when the reflector voltage reaches the value at which the next mode starts up. Correspondingly, from the second equation of (V.26), the relation between frequency and θ will take the form of a tangent curve. This is the relation describing the electronic tuning of the oscillator. We notice that, the smaller the external Q, or the tighter the coupling to the load, the greater is the g value corresponding to a matched load. For a very tight coupling, the vertical line in the figure below on this page may well be so far to the right that it does not intersect the spiral at all, for lower modes. These modes then do not operate with tight coupling. We should realize that not only does this set a limit on the lowest mode which will operate, but there is also a limit set on the highest mode, by the condition that the reflector voltage V_r must clearly be greater than the beam voltage V_0, or the electrons will strike the reflector and be absorbed, rather than being reflected. From (V.6), this means that the maximum value of θ is definitely determined, by the condition

$$Q_{\max}=\frac{2m}{eV_0}\omega[2(e/m)V_0]^{\frac{1}{2}}d_r = 2\sqrt{2}\left(\frac{m}{e}\right)^{\frac{1}{2}}\frac{\omega d_r}{(V_0)^{\frac{1}{2}}}, \quad \text{(V.27)}$$

showing that the lower the beam voltage V_0, the higher is the maximum usable mode.

In addition to studying the operation of the oscillator when looking into a matched load, we may study the operation as a function of load, with fixed reflector voltage. This leads us to the Rieke diagram, as in the preceding chapter. There is nothing unusual about the individual diagrams, but it is interesting to see how they change from one reflector voltage to another. The slope of the lines of constant frequency in the admittance plane is the same as the slope of the corresponding radius in Fig. 26, being horizontal at the position of maximum power, and rapidly becoming steeper and steeper as the reflector voltage is varied on either side of this value. We note that G can vary from zero only to a certain maximum value, related to the g value of the small signal spiral; thus operation in the admittance plane is possible only out to a certain horizontal line, and in the reflection coefficient plane in to a certain circle. As the reflector voltage is shifted away from the value for maximum power, over to the edge of the mode, the maximum value of G decreases toward zero, so that the forbidden circle in the reflection coefficient plane grows until it finally includes the whole region of loads with positive resistive components. Outside that region the oscillator cannot operate. Our formulas for g and b continue to be valid, however; the only difference is that g has the opposite sign, so that the tube acts as a non-linear load rather than an oscillator. We may feed power into a reflex klystron whose reflector voltage is in this non-operating region, and may measure its input impedance. From Eq. (III.81) we see that the electronic term (which may be rewritten by (III.85)) has the effect of modifying the apparent loaded Q, and resonant frequency of the cavity. The term $g+jb$ has such a sign that in the non-oscillatory region, where the electrons act as a load, they increase the value of $1/Q_L$, or add to the loading; in actual cases they reduce the loaded Q to a low value. As the reflector voltage is varied toward the operating range, however, the effect of the electrons on the loading decreases, becomes zero when $g=0$, then changes sign, and helps to cancel the losses resulting from the unloaded Q of the cavity. As the operating range is reached, the apparent Q of the cavity increases without limit, until finally oscillation occurs. Along with this change of the apparent Q with reflector voltage, there is also a change in the apparent resonant frequency of the cavity, as we can at once compute. If only a small amount of power is being fed into the cavity to make the impedance measurement, we may assume that we have a

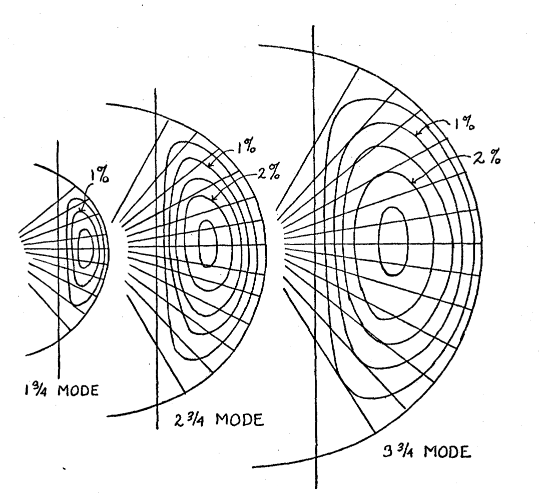

FIG. 26. Efficiency contours in admittance plane, for reflex klystron.

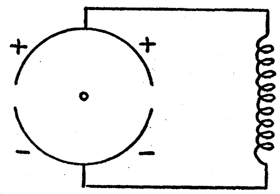

FIG. 27. Schematic diagram of two-anode magnetron.

small signal, and may compute the value of b, and hence of the frequency, from the small signal values (V.23). We have then an electronic means of tuning a resonant cavity, the non-operating klystron. An electronic reactor, a reactance which can be tuned by electrical means, is a very useful device, and in a sense the reflex klystron in its non-operating range forms such a reactor. On account of the very large losses which are associated with the g, however, this use as a reactor is not of practical importance. Other types of tubes can be designed, however, in which there is an electrically controlled reactance, without accompanying large resistive losses.

4. The Magnetron

The multi-segment magnetron oscillator is a much more complicated problem than the reflex klystron, for two principal reasons. First, its oscillating circuit, or resonant cavity, is of a more complicated shape, and it is considerably harder to make approximately correct theories explaining its resonant modes. Furthermore, there is a difficulty arising in the circuit, on account of other resonant modes than the desired one, which come at nearly the same frequency, unless special means are taken to remove them. The second reason for the complication of the magnetron is the electronic motion. The presence of the magnetic field makes a one-dimensional treatment, such as we used for the klystron, impossible, and the existence of large space charge makes any analytical treatment of the motion a very difficult thing. For these reasons, we shall not be able to give a very complete treatment of the theory of the magnetron. Nevertheless we shall be able to go far enough to indicate the reasons for the high efficiency and high power characteristic of this type of oscillator.

The magnetron oscillator, in its present form, consists of a cathode and anode in the form of concentric cylinders, with a constant magnetic field along the axis of the cylinders. The anode, which is outside the cathode, is split into an even number of segments, say N, and forms part of a resonant cavity such that, in the resonant mode in which the magnetron is operated, successive segments are positively and negatively charged. The electrons move, after emission from the cathode, under the action of the magnetic field; an impressed d.c. electric field accelerating them from cathode to anode; and the r-f field between oppositely charged segments of the anode. This r-f field leads to the r-f voltage V which appears in our theory, as in the preceding chapter. As a result of the combination of these fields, the electrons move in a complicated way which we shall describe. They eventually reach the anode, but not with the kinetic energy which they would have acquired if they had fallen directly from cathode to anode under the d.c. difference of potential. Instead, on account of the interaction with the magnetic and r-f fields, they have very small kinetic energy on reaching the anode, so that they dissipate only a small fraction of the input power at the anode. The rest of the input power goes into sustaining the r-f oscillation, and is the electronic power, P_{el}, about which we have previously spoken. We can compute the r-f current amplitude i, which we need in the theory; it is here not at all clear at first sight how we are to find this current, and it is necessary to go back to the definition in terms of the integral $\int \mathbf{J}\cdot\mathbf{E}_a dv$ to find how to compute it. When we find i as a function of V, we find that the general situation is similar to that of the preceding chapter, so that we can compute output power and frequency as functions of load as in that chapter. We also can consider the dependence of power output on the d.c. parameters, the d.c. voltage between cathode and anode, the magnetic field, and the d.c. current which flows. We shall now give a short discussion of the nature of the cavity resonator which produces the r-f oscillations in the magnetron, and shall then consider the electronic motions which lead to its operation.

5. The Resonant Circuit of the Magnetron

For a good many years the split anode magnetron, having an anode of two segments, has been used as an oscillator. Treated from the standpoint of lumped constants, the two anode segments, charged to opposite potentials, form a condenser, and they must be connected by an inductance, as shown in Fig. 27, to make them into a resonant circuit. To increase the power of such an oscillator, it is natural to increase the number of circuits, having many anode segments. By connecting each pair of adjacent segments by an inductance, such a system can be made to resonate. The geometry then allows a large cylindrical cathode, instead of the linear cathode of the split anode magnetron, and this permits the flow of a large anode current, with consequent high power. If the magnetron is to be used in the microwave range, it is natural to use resonant cavities instead of lumped constants for the inductances, and the problem of heat dissipation requires that the anode be made of solid metal. The simplest structure embodying these principles is shown in Fig. 28, and was the type of magnetron which, at the beginning of the war, first showed the possibility of really large power generating capacity. The larger part of the capacity between adjacent anode segments is here concentrated in a slot between the segments, and a hole in a solid copper block forms the inductances. Magnetic lines of force from the inductances thread through the spaces above and below the block, to complete the magnetic circuit.

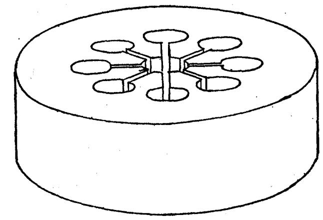

FIG. 28. Anode of multi-segment magnetron.

The problem of the nature of the field in such a cavity may be divided into two parts: the field in the interaction space (that is, the cylindrical region between cathode and anode, in which the electronic discharge is located) and the field in the separate hole and slot resonant elements. The first of these problems can be handled by solving Maxwell's equations in cylindrical coordinates. We may get an approximate physical idea of the problem, however, by imagining the anode and cathode flattened out into planes, as shown in Fig. 29. Then, for the mode in which we wish to operate, the electric lines of force will run as in Fig. 29: the field will be the fringing field of the slots, regarded as condensers. The field will be periodic along the direction parallel to the anode and cathode faces (the z direction in the figure), the effective wave-length being twice the spacing of the segments. This, in actual cases, is much less than the free space wave-length, so that the effective velocity of propagation of this wave along the z axis is much less than the velocity of light. We may regard the field drawn in the figure as a standing wave, a superposition of two traveling waves moving in opposite directions along the $\pm z$ axis. We can show in a very simple manner that if the effective wave-length along z is less than the free space wave-length, the field must fall off exponentially as we go away from the anode; this expresses the way in which the fringing field falls off with distance. To see this most simply, we may regard the interaction space as a wave guide, with propagation along the z axis. Using the expression (II.9) for the relation between free space wave-length, guide wave-length, and cut-off wave-length, we see that if the guide wave-length is smaller than the free space wave-length, as it is here, the cut-off wave-length must be imaginary, corresponding to an exponential rather than sinusoidal variation of the field at right angles to the direction of propagation. Such a variation of course cannot satisfy the boundary conditions in a closed wave guide, for if the field increases exponentially as we approach the anode surface, there will be a tangential electric field at that surface, which would be impossible if we had a conducting surface for the anode. In our case, however, the anode surface is broken by the slots leading to the resonators, and there can be a tangential component of electric field across imaginary surfaces closing these slots. We can, in fact, get a simple and fairly accurate approximation to the actual solution by computing the ratio of tangential E to tangential H, or the impedance, along the anode surface, from the solution holding in the interaction space, and equating this to the corresponding input impedance of the resonant cavities.

For a wave traveling in the $+z$ direction in

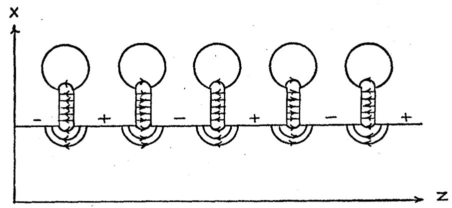

FIG. 29. Electric lines of force in linear magnetron.

the interaction space, we find easily that a solution of Maxwell's equations is

$$E_x = Ae^{\gamma x}e^{j(\omega t-\beta z)}, \quad E_y=0, \quad E_z=-j\frac{\gamma}{\beta}E_x,$$

$$H_x=H_z=0, \quad H_y=\left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}}\frac{\lambda_g}{\lambda_0}E_x,$$

where

$$\omega^2/c^2=\beta^2-\gamma^2, \quad \omega/c=2\pi/\lambda_0, \quad \beta=2\pi/\lambda_g. \quad \text{(V.28)}$$

The standing wave formed by superposing two traveling waves like (V.28) in opposite directions has much the form of that shown in Fig. 29. At the cathode, the field is small, if it is large at the anode, on account of the factor $e^{\gamma x}$, which increases rapidly as we approach the anode; by superposing a similar wave in $e^{-\gamma x}$, we can make the tangential field exactly zero at the cathode, without making appreciable change in the field near the anode. At the anode, there is clearly a tangential component of E, E_z. The wave could be propagated in a guide, if the anode surface had a surface impedance, given by the ratio $-E_z/H_y$, or

$$-E_z/H_y=j(\mu_0/\epsilon_0)^{\frac{1}{2}}[(\lambda_0^2/\lambda_g^2)-1]^{\frac{1}{2}}, \quad \text{(V.29)}$$

as we find easily from (V.28). The series of slots, with their attached resonators, can simulate a surface with this impedance. Looking into one of the slots, a single resonator will have an input impedance which may be written

$$Z_r=Z_0\sum_a \frac{1/Q_{\text{ext},a}}{j[(\omega/\omega_a)-(\omega_a/\omega)]}, \quad \text{(V.30)}$$

as we have seen in previous chapters, where Z_0 is the characteristic impedance of the slot, regarded as a transmission line. If we now have resonators spaced a distance D apart, and if each slot has a width d, we shall find in the anode surface that actually there is a tangential component of E in each slot, while there is no tangential component along the metallic segment. The average field is then d/D times the field in a slot. The current, and hence the magnetic field, however, are continuous along the face of the anode, so that they are not affected by the slots; current flows along an anode segment, into a resonator, out again, and along the next segment, as if the slot were not there. Thus the average impedance of the surface with the slots is $(d/D)Z_r$ (where to make this expression comparable to (V.29) we must assume that the height of the anode is unity; if it is not, it is simple to correct for it). We may then equate this quantity with the value of (V.29). Writing $Z_r=jX_r$, this gives easily

$$\frac{1}{\lambda_g}=\frac{1}{\lambda_0}\{1+(X_nd/D)^2(\epsilon_0/\mu_0)\}^{\frac{1}{2}}. \quad \text{(V.31)}$$

Remembering that the input reactance X_r of a resonator acts like an inductance at long wavelengths, or is proportional to the frequency, we see that, for large values of λ_0, (V.31) shows us that $\lambda_g=\lambda_0$. As the frequency increases, however, and we approach the first resonant frequency of the resonator, X_r increases to infinity, so that $1/\lambda_g$ becomes infinite. It is this first resonance which concerns us at present, so that we need not consider further resonances. We may then easily plot $1/\lambda_g$ as a function of $1/\lambda_0$. For our purposes, as we shall see presently, it is better to plot $1/\lambda_0$ as a function of $1/\lambda_g$. We show this function in Fig. 30. It is clear that as the frequency approaches the resonant frequency, and $1/\lambda_g$ becomes large, or the guide wave-length becomes small, we approach the situation which is actually present in the interaction space of the magnetron.

A formula of the type of (V.31) takes no account of the fact that the anode is made of segments of finite size, and is a periodic structure. When we consider this fact, the theory becomes much more involved. We must represent the field in the interaction space, not by a single wave like (V.28), but by a superposition of an infinite number of waves, satisfying certain periodicity relations, so that a superposition of them, with appropriate coefficients, can actually satisfy the boundary conditions at the anode surface, with zero tangential E along the segments. Carrying out such a calculation, we find two differences in the results, one minor, the other fundamental. The minor result is a small change in the curve of $1/\lambda_0$ as a function of $1/\lambda_g$, for all values of $1/\lambda_g$. The fundamental result is that the curve now becomes periodic in $1/\lambda_g$. We find that there is a minimum value of the guide wave-length for which we can have a solution: twice the distance between slots, or $2D$, using our notation. If the guide wave-length has this value, we have just the solution shown in Fig. 29. If we try to make the guide wave-length less than this, or $1/\lambda_g$ greater, we find that we merely repeat the solution already found for a smaller value of $1/\lambda_g$. The situation is identical with that met in the theory of the weighted string, or of electric filters composed of a succession of identical four-terminal networks. As a result of this, we find that the true curve of $1/\lambda_0$ as a function of $1/\lambda_g$ has the form shown in Fig. 31, resembling the curve of Fig. 30 for small values of $1/\lambda_g$, but then becoming periodic. The value of $1/\lambda_g$ corresponding to the maximum $1/\lambda_0$, or the maximum frequency, is then the type of oscillation in which we are interested.

In the solution of the type shown, we can compute the frequency for a wave of arbitrary guide wave-length. The fact that the curve connecting frequency with $1/\lambda_g$ is not a straight line shows that there is dispersion; the velocity of propagation is a function of frequency. For a linear structure, like that of Fig. 29, any wave-length and any frequency would be possible. For the cylindrical magnetron structure, as shown in Fig. 28, however, the situation is quite different. The anode now closes on itself, and is of finite length; since the field must be continuous in going completely around the anode, we see that the circumference must be a whole number of wave-lengths. This can be automatically handled by using the correct solution of the problem, in terms of Bessel's functions, but we can treat it approximately, from our linear model, merely by demanding that the circumference $2\pi R$, divided by the wave-length λ_g, should be an integer n. That is, we have,

$$1/\lambda_g = n/2\pi R. \qquad \text{(V.32)}$$

In other words, we do not have all values of $1/\lambda_g$ allowed, but only a set of equally spaced discrete values. Such values are shown in Fig. 31, for the case of eight segments. For this case, $2\pi R = 8D$, where D is the width of a segment. For the maximum value of n, we have $1/2D = n/8D$, or $n = 4$; in general, as we see from this example, the maximum value of n is $N/2$. We now see from the figure that each value of n gives a different frequency. Thus we have made an approximate calculation of several of the resonant frequencies of the cavity. Experimental measurement, or more accurate theory, shows that this type of theory is qualitatively correct. It is now obvious from the curve that the modes $n = 3$ and $n = 4$ lie close together. Since it is $n = 4$ in which we wish to operate, this means the presence of a disturbing mode, and for this reason this type of anode is not very satisfactory for actual magnetron operation.

It would take too long to go into the theory of the various means which have been used for separating the resonant modes of the magnetron anode, so that the desired mode (which is often called the π mode, since in it the phase of each segment differs by π from that of the adjacent segment) shall have no other modes very close to it. The simplest method is called strapping, and may be described in one of its forms as follows. In Fig. 29, we imagine two parallel wires or straps located over the segments of the anode, parallel to the anode surface. One of these straps is connected, by short wires or posts, to all the segments marked + in the figure, and the other strap to all segments marked −. It is then clear that, in the mode shown, one strap will be positively charged, the other negatively. Thus the strap will act like an added capacity in shunt with the capacities of the slots, hence increasing the capacity, and decreasing the resonant frequency of the π mode. For the other modes,

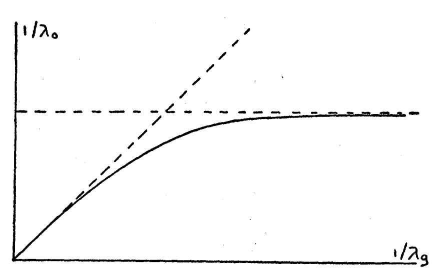

FIG. 30. Frequency as function of reciprocal wave-length, for linear magnetron.

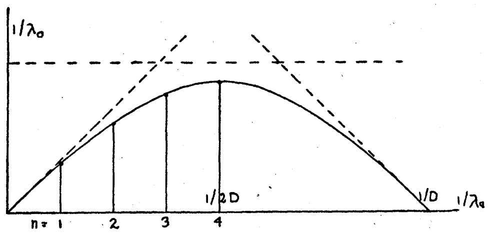

FIG. 31. Frequency as function of reciprocal wave-length, for eight segment magnetron.

however, each strap is charged positively over part of its length, negatively over the rest, and thus the system of straps is less effective as a condenser. When we examine the theory, or measure the position of the modes experimentally, we find that the displacement of modes to lower frequency, or longer wave-length, decreases regularly as we go from the π mode to modes of lower n value. This effect is much greater than the mode separations present in an unstrapped magnetron, which are shown in Fig. 31, and which are in the opposite direction. Thus the order of modes is reversed, and in a strapped magnetron the π mode has the lowest frequency, or longest wave-length, the frequencies of the other modes coming at higher and higher frequencies as n decreases, and being separated far enough from each other so that there is no interference between them.

The reason why other modes are undesirable in a magnetron is mainly that with two modes nearby, the electronic discharge may be unable to decide which of the two modes to operate in, and there may be power emitted in two or more frequencies. It is unlikely that simultaneous operation in two modes is possible; more often, in a pulsed magnetron, either successive pulses are in different modes, or the magnetron shifts mode during a pulse from one mode to another. Either of these phenomena decreases the power in the mode in which operation is desired. These difficulties of modes have been among the most troublesome and least understood features of magnetron operation and construction. We have not time to go into them further, and shall assume in our discussion of the electronic motions in the magnetron that the resonant cavity has only one mode, the π mode. The simple theory of Eq. (V.28), and Fig. 29, give a fairly satisfactory description of this mode.

6. Electron Motions in the Magnetron

Now that we have a fairly correct picture of the r-f field inside the interaction space of the magnetron, we can ask how electrons move, in the combined d.c. electric field, along the x axis of Fig. 29, the magnetic field along the y axis, and the r-f field. Here, as in discussing the resonant oscillations, it is convenient to use the linear model with a plane cathode, rather than the actual cylindrical case. As a first step, we consider the motion of an electron in the d.c. fields, without the presence of the r-f field.

An electron of charge $-e$, moving in a constant field of magnitude E accelerating it along the $+x$ axis, and a magnetic induction B along the y axis, has an equation of motion

$$m\ddot{x}=eE+eB\dot{z},\quad m\ddot{y}=0,\quad m\ddot{z}=-eB\dot{x}. \qquad \text{(V.33)}$$

We may satisfy the second by setting $y=0$, and the third by setting $m\dot{z}=-eBx$, which is consistent with the initial condition that the electron starts from rest, or has $\dot{x}=0$, $\dot{z}=0$, when it leaves the cathode, which we take to be $x=0$. Substituting in the first equation, we then have

$$\ddot{x}+\omega_H{}^2x=\frac{e}{m}E,\quad \text{where}\quad \omega_H=\frac{eB}{m}. \qquad \text{(V.34)}$$

Solutions of these equations, satisfying the initial conditions, are

$$x=\frac{mE}{eB^2}(1-\cos\omega_H t),$$
$$z=-\frac{Et}{B}+\frac{mE}{eB^2}\sin\omega_H t, \qquad \text{(V.35)}$$

where t is the time measured from the instant when the electron leaves the cathode. We see that the electron describes a circle, of radius mE/eB^2, with angular velocity ω_H, about a point $x=mE/eB^2$, $z=-Et/B$. We readily see that the resulting path is a cycloid. For we remember that a cycloid is the path made by a point on the rim of a rolling wheel. The wheel is rotating about a point whose height is its radius, and whose linear velocity is the product of the radius and the angular velocity, as our value E/B is the product of (mE/eB^2) and (eB/m). Thus the path is as shown in Fig. 32. The maximum height of the orbit above the cathode is $2mE/eB^2$. If this distance is less than the distance to the anode, the electrons will never reach the anode, and the magnetron will not pass a d.c. current. If however the height is more than this, all electrons leaving the cathode will reach the anode, so that there will be a d.c. current. The value of voltage for a given magnetic field, or magnetic field for a given voltage, for which the electrons just reach the anode are called the cut-off voltage and cut-off magnetic field respectively. The range of values of E and B for which the magnetron can act as an oscillator are those for which the electron would fail to reach the anode in the static case; in fact, ordinarily those for which it fails by a large margin. That is, it acts as an oscillator for magnetic fields large compared to the cut-off value.

With this knowledge of the motion of an electron in the static electric and magnetic fields, it is not hard to see how electrons move in the actual field of the magnetron. In the type of operation which leads to an efficient oscillator, there proves in the first place to be a resonance relation occurring between the average drift velocity of the electrons, on account of the motion of the center of the cycloid, and the velocity of the electromagnetic wave in the interaction space. We have seen in the preceding section that this wave consists of a superposition of two traveling waves, traveling in opposite directions. Each of these waves travels with a velocity less than the normal velocity of light; in fact, in practical cases, much less, perhaps a tenth of the velocity of light or less. The drift velocity of the electron, by (V.35), is E/B. The resonance condition we have mentioned is that these two velocities should agree. That is, there is a linear relation between E and B required for good operation of the magnetron. We can readily find the constant of proportionality. If D is the spacing of the segments, the effective wave-length of the disturbance in the interaction space must be $2D$, so that the ratio of wave-length to the free space wave-length is $2D/\lambda_0$, and this must equal the ratio of the velocity, to the velocity of light. Thus we must have

$$\frac{E}{B}=\frac{2D}{\lambda_0}c. \qquad \text{(V.36)}$$

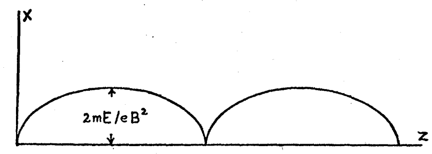

FIG. 32. Cycloidal orbit of electron in magnetic and electric fields.

We can, as a matter of fact, satisfy resonant conditions similar to this for the other modes of the resonant cavity. From (V.32) we see that the wave-length of the disturbance is inversely proportional to n, the index of the mode; thus the velocity varies in the same way, and we have

$$\frac{E}{B}=\frac{n_{\max}}{n}\frac{2D}{\lambda_0}c, \qquad \text{(V.37)}$$

where $n_{\max}$ is the maximum value of n, which we have for the π mode. This relation is not quite as simple as it seems, for we must remember that λ_0, the free space wave-length connected with the frequency of oscillation, is itself a function of n. For the unstrapped tube, as shown in Fig. 31, $1/\lambda_0$ is proportional to n for small n values, so that E/B is constant for these values, and in fact is approximately equal to the velocity of light; it is only for the higher n values that $1/\lambda_0$ becomes small enough, and $n\lambda_0$ big enough, to reduce the velocity well below the velocity of light. On the other hand, with strapped tubes, we have seen that λ_0 becomes considerably greater than for unstrapped tubes, so that the velocity becomes considerably less than the velocity of light.

We have stated that the resonance condition must be satisfied, according to which the drift velocity of the electrons equals the velocity of the traveling wave. Now we must ask why this is so. If the condition is satisfied, then as the electron moves along, the r-f field which travels along with it will appear to it to be a constant field. Then we can immediately find its motion in this field. We need merely take the vector sum of the d.c. field, and of the r-f field which appears to be constant, and find the motion of an electron in this constant field. If the phase of the electron is such that the r-f field following along with it is in the x direction, or from the anode to the cathode or *vice versa*, then this r-f field will merely add in magnitude to the d.c.

field, and the result will be that the drift velocity of the electron will be increased or decreased. On the other hand, if the r-f field is along z, or at right angles to the d.c. field, then the vector sum of the r-f and d.c. field will be in a different direction from the d.c. field. The motion will then be similar to the cycloid of Fig. 32, but the direction will be different. Depending on the phase, the path will either carry the electron down into the cathode, in which case it will strike the cathode after its first revolution, and be lost; or it will carry it away from the cathode, in which case the electron will move in a diagonal path toward the anode. It is electrons of this type which act as generators. For they move from cathode to anode, and yet at the anode they do not have a very large kinetic energy. Since they would have acquired a large kinetic energy falling from cathode to anode in the absence of an r-f field, it must be that the remaining energy was used up in working against the r-f field, and therefore must have been delivered to the resonant circuit as electronic power. Thus it is not surprising that the magnetron can have high efficiency.

Let us first consider electrons leaving the cathode in different phases, and see how they act. Referring to Fig. 29, we see first that those electrons for which the r-f field is in the direction of the d.c. electric field, and which therefore drift faster than the velocity of the field, will then catch up with the field, and enter a region where the r-f force is opposite to the drift velocity. This combines with the d.c. force to give a net force such that the electron drifts toward the anode, in the manner described above. On the other hand, those electrons which have the r-f field opposite to the direction of the d.c. field will tend to be slowed down, and will likewise enter a region where the r-f force is opposite to the drift velocity. There is, in other words, a tendency toward bunching of the electrons, into the phase in which they can operate as efficient generators of power. The only electrons which do not work in this way are those which start out with a component of r-f force in the direction of their velocity. They will immediately start to drift further toward the cathode, and will be lost at the end of their first cycle. They will dissipate some power at the cathode; for the r-f force will have worked on them during the one cycle of their motion, and they will have accumulated some kinetic energy, instead of returning to the cathode with exactly no kinetic energy, as in the static case of Fig. 32. This dissipated power is far less, however, than that delivered to the r-f field by all the other electrons, which finally reach the anode.

We can make a simple approximate calculation of the efficiency to be expected from the magnetron, as a function of the d.c. voltage and magnetic field. Let the distance from cathode to anode be X. Then an electron gains potential energy of eEX, or of eV_0, where $V_0=EX$ is the d.c. voltage, in falling from cathode to anode. Part of this appears as kinetic energy of the electron, which is dissipated in the collision with the anode, and lost. The rest is delivered to the r-f field. We can easily find the average kinetic energy of the electrons. The electrons are most likely to strike the anode when at the top of their cycloidal path. At such a point, we can find easily, from (V.35), that the kinetic energy is $2m(E/B)^2$. Thus the energy put into an electron by the d.c. field is eV_0, the energy dissipated is $2m(E/B)^2$, and the useful energy delivered to the r-f circuit is the difference of these quantities. The electronic efficiency, or ratio of electronic power to input power, is thus $(eV_0-2m(E/B)^2)/eV_0$. This may be rewritten in the form

$$\eta_{el}=1-(2m/e)(E/B^2X). \qquad \text{(V.38)}$$

We may put this in a convenient form in terms of the cut-off voltage and magnetic field. At cut-off, the maximum height of the cycloid, $2mE/eB^2$, equals the distance X from anode to cathode. Thus we have

$$(2m/e)(E_c/B_c^2)=X, \qquad \text{(V.39)}$$

where E_c, B_c are cut-off electric and magnetic fields. Substituting this value for X in (V.38), and using the fact that the cut-off voltage is E_cX, we have

$$\eta_{el}=1-[(V/B^2)/(V_c/B_c^2)]. \qquad \text{(V.40)}$$

This very simple formula is surprisingly accurate for giving the general form of the electronic efficiency of a magnetron. It shows that the efficiency is zero at cut-off, but increases as the magnetic field is increased or the voltage decreased, so that we are farther and farther from cut-off. Furthermore, no upper limit is indicated for the efficiency; it appears from (V.40) that it can become arbitrarily close to unity, for sufficiently high magnetic fields. As far as is known experimentally, there is nothing to contradict this feature of the theory. Magnetrons have been operated with electronic efficiencies in the neighborhood of 90 percent, in striking contrast to reflex klystrons, in which efficiencies of a few percent are common.

7. Operating Characteristics of the Magnetron

We have seen in the preceding section something of the type of electronic motion in the magnetron, and the physical reason for its high efficiency. To proceed further, however, we wish to find the r-f current as a function of r-f voltage, to find the quantities g and b, and to discuss operation as a function of load. Furthermore, we wish to understand the particular sort of operating curves convenient for practical discussion of magnetron operation, in which we are concerned with the relations between d.c. current, voltage, and magnetic field. The ordinary experimental test of a magnetron is conducted with a matched output, and consists of an observation of the relation between d.c. current and voltage at different values of magnetic field as a parameter. The power, or efficiency, are observed as functions of current and voltage. The commonest type of plot, the performance chart, has d.c. current as abscissa, d.c. voltage as ordinate, and consists of lines of constant magnetic field and of constant power or efficiency, in such a plane. We shall now try to understand the theory underlying such a performance chart.

First let us inquire what is the d.c. current flowing to the anode, with a given d.c. voltage, magnetic field, and r-f voltage. It is clear from the discussion of the preceding pages that the r-f voltage makes electrons drift across from cathode to anode, even when the magnetic field is beyond cut-off, or when the d.c. electric field is so small that it by itself would not carry electrons to the anode. Furthermore, the tilt of the orbit, and hence the drift velocity of the electrons, will be roughly proportional to the r-f voltage. The d.c. current will then be proportional to the density of space charge times the drift velocity. If the discharge is space charge limited, a more elaborate discussion than we can give here indicates that the total space charge density is roughly independent of the r-f voltage. Thus we conclude that the d.c. current will be proportional to the r-f voltage. Furthermore, the nearer the d.c. voltage is to cut-off, the larger will be the d.c. current. This is partly on account of the natural increase of space charge density as the voltage increases, since the space charge limited current would be proportional to the 3/2 power of the voltage in the absence of a magnetic field. It is partly also on account of the fact that as we approach cut-off, the cycloids become nearly as large as the distance between cathode and anode, and a small drift of the cycloid toward the anode will cause the electron to strike the anode. Using these principles, we may then deduce that the curves of d.c. current as a function of d.c. voltage, for a variety of r-f voltages, and a given magnetic field, have the form shown in Fig. 33. At zero r-f voltage, no current theoretically will flow until the voltage reaches its cut-off value. At this point the current will suddenly jump to a value determined by space charge theory, as influenced by the magnetic field. It will then increase rapidly with increasing voltage, not just according to the 3/2 power law given by space charge theory, but somewhat more rapidly, as one can find by considering the effect of magnetic field. With a finite r-f voltage, current will begin to flow at a smaller d.c. voltage, and will increase rapidly with d.c. voltage. The exact form of the curves is not known either from detailed theory or from experiment, but it presumably is much as shown in the figure. We can see clearly from the figure that at constant d.c. voltage the current increases with r-f voltage, as it should.

A graph like that above gives only part of the information we need to interpret magnetron operation. We need also a curve giving r-f current as a function of r-f voltage. By knowing that, and using the principles of the preceding work, we can find the r-f admittance $g+jb$ coming from the electronic discharge. Then if we

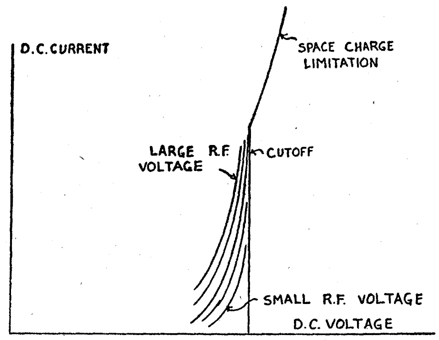

FIG. 33. D.c. current as function of d.c. voltage, for magnetron with various r-f voltages.

know the load, we can find the r-f voltage. Going back to the curve of Fig. 33, this locates us on a definite curve of d.c. current *vs.* d.c. voltage, so that if we know one of these quantities, we can find the other. We must then consider the r-f current flowing in the magnetron. This is less simple to understand than in the reflex klystron; we must really go back to the fundamental formula involving $\int \mathbf{J}\cdot\mathbf{E}_a dv$ to find what it is. The quantity E_a is the field in the mode we are interested in; that is, it is the fringing field as shown in Fig. 29. The current density in the neighborhood of this fringing field, as we have seen in our earlier discussion, is located in bunches, in the part of the field opposing the direction of drift of the electrons. The component of current density in the direction of this field is parallel to z. That is, it is proportional to the product of component of drift velocity in that direction (which is relatively independent of r-f voltage) and charge density (which is also roughly independent of r-f voltage). Thus it arises that there can be a large r-f current even for very small r-f voltages. As long as the r-f voltage is large enough to produce bunching, which may require only a relatively few volts, the bunches will have as much charge density, and will be moving with the same drift velocity, as if the r-f voltage had been much greater. This is the situation which was mentioned in earlier sections as the peculiar property of the current in the magnetron, in which a small r-f voltage almost discontinuously produces a large r-f current. As the r-f voltage increases, however, and becomes comparable to the d.c. voltage, the electrons tend more and more to drift straight across to the anode, the angle of tilt of the orbit getting greater and greater. They are then less favorably inclined to the r-f field, and as a matter of fact when calculations are made the r-f current proves to decrease, apparently approaching zero for r-f voltages sufficiently great. This applies only to the component of r-f current in phase with the r-f field; the component out of phase, the reactive component, is large at all values of r-f voltage, and results in a reactive tuning of the magnetron. The problem of determining the Rieke diagram of a magnetron, then, is not greatly different from the other cases we have previously considered, and we need not give a separate discussion for it.

The curve of r-f current *vs.* r-f voltage of course will depend on the d.c. parameters. As the d.c. voltage increases, the space charge density will increase, and the r-f current will increase, as does the d.c. current. Thus we find that curves of r-f current *vs.* r-f voltage, for a set of d.c. voltages or of d.c. currents, have the appearance of the curve of Fig. 34. We can read off from these curves the decrease of r-f current with increasing r-f voltage at a constant d.c. voltage, the r-f current approaching a constant value at small r-f voltages; and also the rapid decrease of r-f current, at a given r-f voltage, with decrease of d.c. voltage. Furthermore, we see the corresponding decrease of d.c. current, and as a result see that the lines of r-f current as a function of r-f voltage, at constant d.c. current, are much more sloped than those at constant d.c. voltage. The ordinary Rieke diagrams of magnetrons are usually taken at constant d.c. current, and it is accordingly these strongly sloping curves which must be used in discussing them.

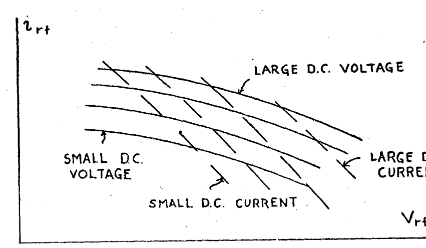

FIG. 34. R-f current as function of r-f voltage, for magnetron with various d.c. voltages and currents.

From these curves of r-f current as a function of r-f voltage, we can now find the r-f voltage with which the tube will operate, with a given load and given d.c. conditions. If we know the load admittance, then by methods that we have often used we can find the ratio $g+jb$ of r-f current to r-f voltage. The curves of Fig. 34 show the component of r-f current in phase with the voltage; thus a line of constant g is a straight line through the origin. Knowing the load admittance, we draw such a straight line. Its intersection with the line of constant d.c. voltage, or constant d.c. current, appropriate to the circumstances, shows the r-f voltage and current at which the tube operates. One corollary of this is clear. For a given load, we are always operating on a given straight line through the origin. We then see, from the way the curves are arranged, that as either the d.c. voltage or the d.c. current increases, the r-f voltage and r-f current will increase. Since the electronic power is proportional to the product of r-f current and r-f voltage, the power will also increase with increase of d.c. voltage or current. All these considerations assume that the magnetic field is constant during the discussion.

We can now return to Fig. 33, and find the curve of d.c. current *vs.* d.c. voltage, not for constant r-f voltage, but for constant load, as in an ordinary test of performance. Since the d.c. current increases rapidly with r-f voltage, the curve will cut across the curves of constant r-f voltage, falling at higher and higher r-f voltage, and higher d.c. voltage, as the d.c. current increases. That is, the curves of constant load will look as in Fig. 35, where we give two curves, one for small g, one for large g (that is, the first one is for a loose coupling, for which the r-f voltage is large, and the second for tight coupling). We see that the increase of g decreases the r-f voltage at constant d.c. current, and hence increases the d.c. voltage. A curve of constant load, and constant magnetic field, is then, as this figure shows, approximately a straight line, starting at small d.c. currents from a value of d.c. voltage somewhat below cut-off, and with the d.c. voltage increasing slightly with increase of d.c. current. Of course, as the magnetic field is changed, the d.c. cut-off voltage changes, increasing as the magnetic field increases, so that the line representing operation at constant load shifts to higher d.c. voltage with increase of magnetic field. We thus see the origin of the performance chart of the magnetron, which takes the form shown in Fig. 36. It is usual to interchange the axes, plotting d.c. voltage as a function of d.c. current, and drawing lines of constant magnetic field (and constant load) on the chart. In addition, the lines of constant power are plotted on the chart (and sometimes also the lines of constant efficiency). We can find the power from the information already stated. Since we have seen that, at constant magnetic field and load, the power continuously increases with increasing d.c. voltage or current, the lines have the general form shown in the figure. To find the efficiency, we must divide the power (which will be the electronic power times the circuit efficiency) by the input power, the product of the d.c. current and

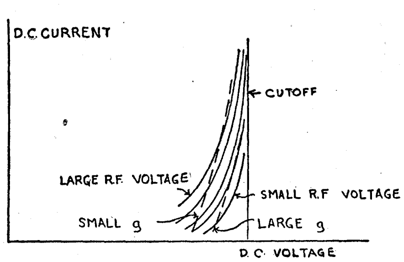

FIG. 35. D.c. current as function of d.c. voltage, for magnetron with various load conductances.

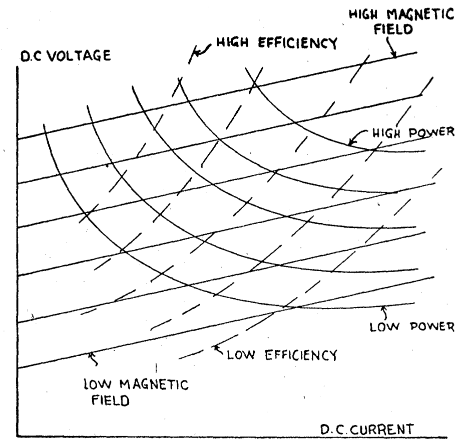

FIG. 36. Magnetron performance chart.

d.c. voltage. When we do this, we find that the efficiency has its maximum values at low d.c. currents, but high voltages, as shown in the figure. The curves shown above agree with the actual performance charts of magnetrons, except in the very low current region. In that range, as we see from the preceding page, the r-f voltage is very small, and many disturbing features can come into the operation, in particular the interference of other modes. These interferences have the effect of decreasing the power and efficiency, so that as a rule the efficiency, rather than having its maximum value for low current, is low at that point, rapidly increases as the current increases, and then decreases again with still further increase of current.