Ultrahigh Frequency Engineering

THOMAS L. MARTIN, JR.
ASSOCIATE PROFESSOR OF ELECTRICAL ENGINEERING
UNIVERSITY OF NEW MEXICO

PRENTICE-HALL, INC.
NEW YORK

PRENTICE-HALL ELECTRICAL ENGINEERING SERIES

W. L. EVERITT, *Editor*

First Printing.....October, 1950
Second Printing.....September, 1953

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

THIS BOOK comprises a general study of the techniques associated with ultrahigh frequency systems and is intended for senior students majoring in electrical engineering or physics. The text deals with the theory and technique of the component parts that make up complete systems, but not with the functioning of such systems as a whole. The book provides a firm, fundamental background in the general subject of ultrahigh frequency engineering to qualify the student for such new fields of electronic engineering as radar, facsimile, television, blindlanding systems for aircraft, ionosphere measurements, pulse-time modulation, telemetering, and electronic computing. The book is not designed to prepare the student for any one of these specialized fields, but to acquaint him with the general principles that apply to all of them. The extensive applications of ultrahigh frequency engineering, and the constant enlargement of these applications, precludes a complete and exhaustive treatment of the subject in one volume.

Six general classifications of subject material are discussed in *Ultrahigh Frequency Engineering:*

1. Generation and synchronization of special waveforms.
2. Amplification for UHF systems.
3. UHF transmission systems.
4. UHF circuit elements.
5. UHF oscillators.
6. Propagation of signals.

In general, the fundamental theory is developed within each major classification, followed by selected generalized examples and discussions. The theory is presented with the idea of developing *concept;* the examples are used to illustrate *technique* of analysis. In every case a conscientious effort is made to stress similarity.

Mathematics through calculus is assumed, as well as the usual introductory courses in electron tubes and circuits. Differential equations, both partial and total, are used in the text, but all steps

involving their use have been carried through completely and accompanied by considerable verbal explanation. Consequently, their use should not occasion any serious difficulty for the student who is inexperienced in this branch of mathematics.

The author is indebted to Professor H. D. Harris of the Electrical Engineering Department, Rensselaer Polytechnic Institute, under whose guidance this text was initiated, to his wife, Hélène H. Martin, whose encouragement and forbearance made this book possible, and to the *Bell System Technical Journal*, April, 1949, for the inspiration of Figure 10.20. The author is also indebted to Dean W. L. Everitt, of the College of Engineering, University of Illinois, for many helpful suggestions in the final preparation of the manuscript.

Thomas L. Martin, Jr.

Albuquerque, New Mexico

CONTENTS

CHAPTER 1

WAVE-SHAPING CIRCUITS

PROPER OPERATION of many ultrahigh frequency (UHF) systems depends upon the development and control of special waveforms that are not ordinarily available from some simple circuit. Generally, though not always, such systems are used in conjunction with the tremendously versatile cathode ray tube, as in radar, television, and blind landing systems, to name only a few. The requirements placed upon the motion of the electron beam in a cathode ray tube employed in any of these systems are generally quite severe, both in the complexity of the motion itself, and in the precise timing, or synchronization, which must be provided to maintain proper sequential operation within a single device, or between a group of devices.

The circuits required to produce these complicated and unusual waveforms, with the necessary precision of form and relative time phase, are often exceedingly complex. It is common for wave-shaping channels to contain a dozen or more tubes, and systems having fifty or even a hundred tubes are not considered to be unduly remarkable.

On the surface, such networks appear to be terribly complicated, but they have the intrinsic property of being derived from elementary constituent circuits that are usually quite simple in operation and design. The situation is analogous to a sculptor viewing his completed statue. The lines, planes, and curves appear very complex, but each was constructed by chipping a little off here and there so that the final statue is constructed through a series of shaping operations. The know-how required in making each individual chip is, of course, essential. The same situation exists in electronic wave shaping. The engineer is required to produce a given waveform from some readily available one, such as a sine wave. He then performs a sequence of shaping operations and "sculpts" the desired signal. Each individual operation is usually confined to a comparatively simple circuit, in most cases consisting of a single tube and a few

allied parts. By representing each constituent circuit which performs a single operation as a block, and then drawing the complete functional block diagram, the original, apparent complexity diminishes considerably.

The numbers and types of circuits which perform shaping operations are legion. There are devices such as clippers, clampers, integrators, peakers, ringing circuits, counting circuits, frequency multipliers and dividers, delay circuits, and so on. Moreover, there are a number of generators of nonsinusoidal waveforms, such as square waves, pulses, sawtooths, trapezoids, hyperbolas, and parabolas, which are produced by various trigger circuits such as multivibrators, relaxation oscillators, blocking oscillators, and pulse-forming lines. The complete list is practically endless. Other components such as rectifiers, filters, phase inverters (amplifiers), as well as the common types of sine wave oscillators, are already familiar to the student through the usual introductory courses.

To cover *all* of these circuits in detail is neither feasible nor desirable. There are a few fundamental principles which can be applied to the bulk of these circuits which show their essential similarities rather than their incidental differences. Thus, the material presented in this and the following chapter is an introduction to a new and fascinating field of electrical engineering.

1.1 The Overdriven Amplifier and Its Equivalent Circuit

Up to this point, the usual electrical engineering curriculum makes extensive use of the familiar and convenient fiction of the equivalent plate circuit theorem, and the analysis of vacuum tube circuits in which small sinusoidal signals are applied to the control grid. A very considerable part of the course of study has been devoted to methods of obtaining faithful reproduction of the input waveform and, consequently, in minimizing and removing possible sources of distortion. This elysium deteriorates rapidly upon the introduction of wave-shaping circuits because every conceivable type of distortion is deliberately introduced to obtain the desired results. The equivalent plate circuit theorem is, in most cases, of no further use in the analysis of circuits of this type.

It is characteristic of most circuits devoted to wave-shaping operations that the function of the tube is restricted to that of a very efficient and quick-acting switch, this method of operation being

produced by the simple expedient of overdriving the tube so that it operates over a very nonlinear portion of its static characteristic, over practically the entire possible range. Proper understanding of this action is absolutely essential to effective study of the more advanced circuits, and will be reviewed in detail in the succeeding paragraphs.

The circuit diagram of a typical triode amplifier is shown in Fig. 1.1, where the signal source E_g is a constant voltage generator. Figure 1.2 shows the load line plotted on the static characteristics of the tube. For any given condition of operation, the voltage across the tube and the current through it are constants. The ratio of

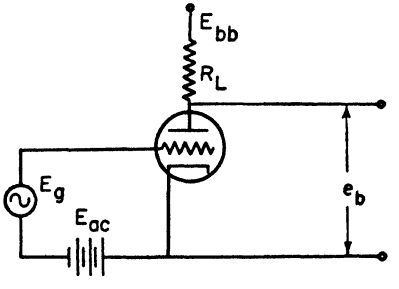


Fig. 1.1. Triode amplifier.

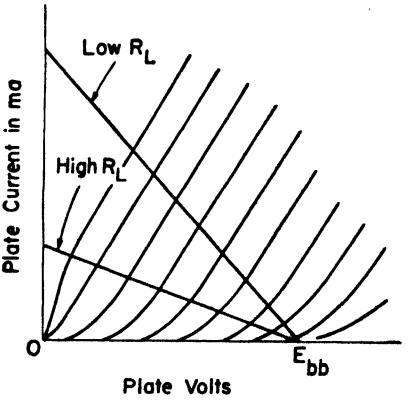


Fig. 1.2. Load line drawn on triode static characteristics.

these two quantities is designated as r_p, and called the static plate resistance. It should be understood that there is no actual resistance in the tube itself, since, if it is properly evacuated, the electrons do not encounter obstructions of any kind in traversing the space between the cathode and the plate. The static plate resistance is simply a measure of the energy required to accelerate the electrons through this space, the energy finally appearing as heat on the plate, due to the release of the ~~electronic kin~~etic energy during impact. An examination of the typical triode characteristics given in Fig. 1.2 shows that the ratio of plate voltage e_b to plate current i_b is far from constant as a function of the grid voltage e_c, varying in a complex manner with the tube electrode potentials and load resistance. Figure 1.3 illustrates the manner in which the static plate resistance

varies with different grid potentials, but with a fixed load R_L and supply voltage E_{bb}.

Inspection of Fig. 1.3 shows that the static plate resistance can vary over a tremendous range of values, but that at any particular operating point, the tube could be replaced by a resistance $\bar{r}_p$. Thus, we can distinguish two limiting cases, regardless of the value of E_{bb} or R_L:

(1) When the tube is cut off, $i_b = 0$, so $\bar{r}_p$ = open circuit..
(2) When the grid voltage is equal to or greater than zero, i_b = is large, so $\bar{r}_p$ = low (almost constant) resistance.

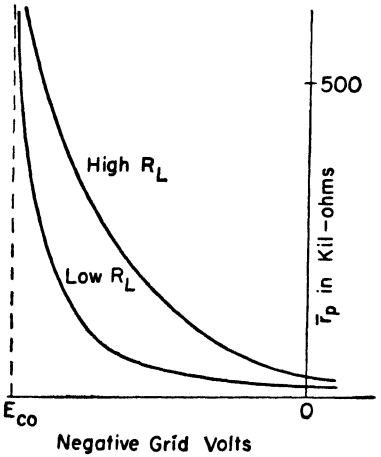


Fig. 1.3. Variation in static plate resistance as a function of grid voltage.

It should be observed that the static plate resistance never reaches zero, but approaches that value asymptotically. In most cases, the lowest value of plate resistance (which corresponds to the highest value of plate current) is determined by the available supply voltage rather than the emission capabilities of the cathode, since there must always be some drop across the tube in order to produce the tube current, and the total drop produced through the tube and load resistance cannot exceed the supply voltage.

Consequently, if we were to apply a signal to the grid of the amplifier in Fig. 1.1, of such magnitude that it exceeded plate current cutoff on the negative swing, and became greater than zero on the positive swing, from the functional standpoint the tube could be replaced by the circuit of Fig. 1.4. When the tube is cut off the switch is in the open position; when the grid is driven positive, the switch closes through a small resistance of the order of 10,000

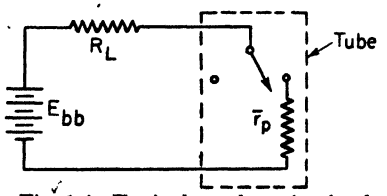


Fig. 1 4. Equivalent plate circuit of the overdriven amplifier.

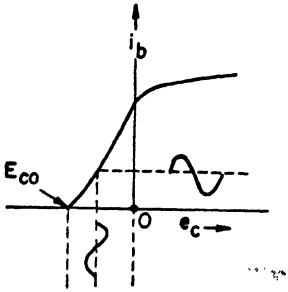


Fig. 1.5(a). Straight amplification.

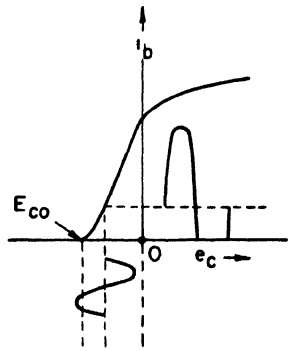


Fig. 1.5(b). Plate current cutoff clipping.

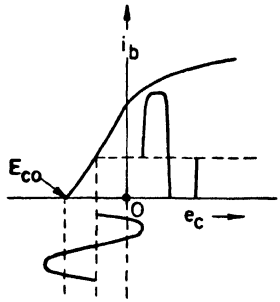


Fig. 1.5(c). Plate current cutoff and saturation clipping.

ohms (for a receiving type of triode). The frequency of the switching action is determined by the nature of the grid signal. In most wave-shaping circuits, Fig. 1.4 is the equivalent plate circuit of the tube.

It is interesting to inquire about the nature of the output waveform in the overdriven operation of an amplifier as discussed in the preceding paragraphs. The operation is best discussed and illustrated by use of the tube transfer characteristics as shown diagrammatically in Fig. 1.5.

Figure 1.5(a) shows conventional amplifier operation where incremental signals are applied over the linear portion of the characteristic. Figure 1.5(b) illustrates the effect of overdriving the tube in the negative region producing what is commonly called "plate current cutoff clipping," since a part of the lower half of the output

current waveform is clipped off due to nonconduction of the tube. Figure 1.5(c) shows plate current cutoff clipping as well as "plate current saturation clipping," the latter being caused by overdriving the tube positively into the region where the static plate resistance (and hence the plate current) is substantially constant. Due to

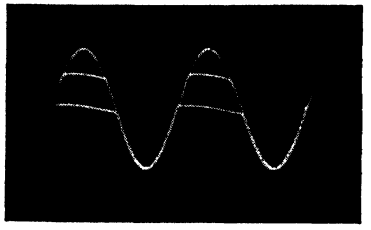


Fig. 1.5(d). Oscilloscope photograph showing the effect of different degrees of saturation clipping.

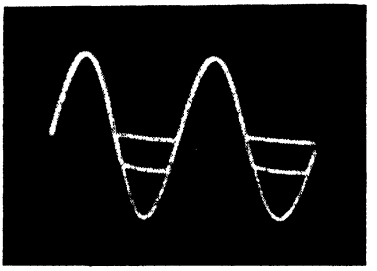


Fig. 1.5(e). Oscilloscope photograph showing different degrees of cutoff clipping.

inversion of the plate voltage relative to the grid voltage, the waveform of the plate voltage will be a mirror image of the plate current.

By way of summation then, we note that:

(1) Overdriven triode amplifiers may be represented by an equivalent plate circuit of the form given in Fig. 1.4.
(2) When the triode is overdriven, clipping of the output waveform occurs as a result of two causes:
   (a) Plate current cutoff.
   (b) Plate current saturation.

1.2 Grid Circuit Clipping and the Equivalent Circuit

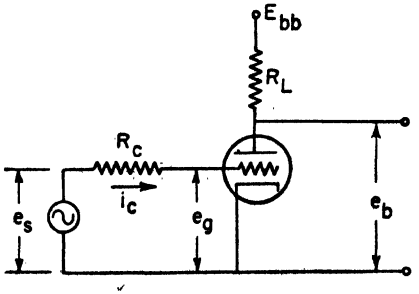


Fig. 1.6. Grid circuit clipping.

The flow of grid current resulting from overdriving the grid of the amplifier discussed in Art. 1.1 in the positive direction may be employed to produce another very effective type of clipping known as "grid circuit clipping." Consider the circuit diagram of Fig. 1.6, which is the same as the triode amplifier of Fig. 1.1 except

that a series resistance has been inserted in the grid circuit. When the signal voltage swings positive, a relatively large amount of grid

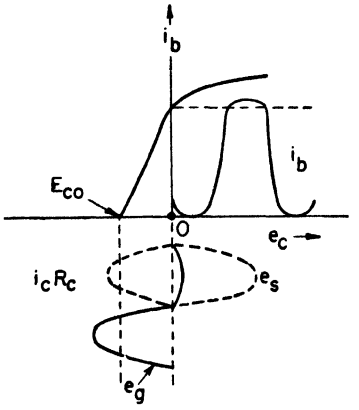


Fig. 1.7(a). Triode transfer characteristic and waveforms.

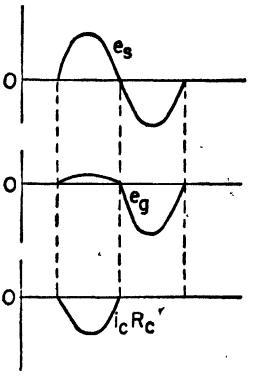


Fig. 1.7(b). Separation of the waveforms shown in Fig. 1.7(a).

current flows, producing a voltage drop in the resistor labelled R_c. This $i_c R_c$ drop nearly cancels the positive swing of the grid, producing a slightly rounded, but otherwise almost square top on the output

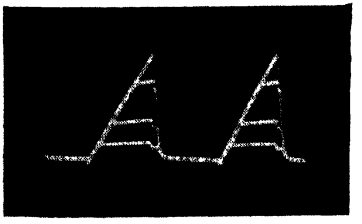


Fig. 1.8(a). Grid circuit clipping when a sawtooth voltage is applied.

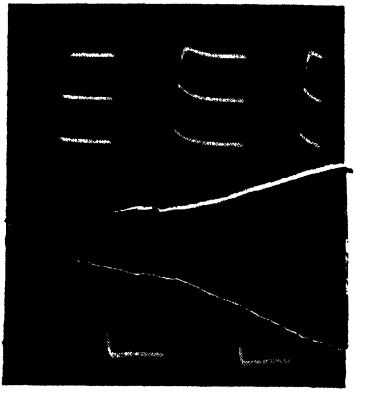


Fig. 1.8(b). Grid circuit clipping, square wave applied. Note the effect of the slight peak at the leading edge of the square wave.

current waveform as shown in Figs. 1.7 and 1.8. Consequently, no matter how positive the signal voltage e_s becomes, the grid-cathode voltage e_c seldom becomes more than a few tenths of a volt positive. In many cases, R_c may be the internal impedance of the signal source and may not appear physically in the circuit.

The mechanism by which the clipping occurs may be more understandable from the development of the equivalent grid circuit. During the time that the grid is positive, the cathode-grid circuit effectively forms a diode, so that a static resistance is present in the circuit, measuring the energy required to accelerate the electrons from cathode to grid. This resistance is usually denoted by $\bar{r}_g$, and is generally about the order of 1000 ohms. On the other hand, R_c is generally made rather large in order to increase the effectiveness of the clipping action. From the equivalent circuit shown in Fig. 1.9, the grid-to-cathode voltage is seen to be

$$e_c = e_s\left(\frac{\bar{r}_g}{\bar{r}_g + R_c}\right)$$

Dividing through by $\bar{r}_g$ yields

$$e_c = \frac{e_s}{1 + (R_c/\bar{r}_g)} \tag{1.1}$$

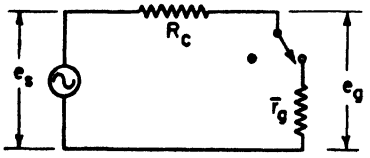


Fig. 1.9. Equivalent grid circuit for grid circuit clipping.

Since the static grid resistance is much less than R_c, then the grid voltage e_c is very much less than the signal voltage e_s. Consequently, as the static grid resistance becomes very small compared to the clipping resistor R_c, the grid voltage approaches very low values, regardless of how positive the excursion of the signal voltage becomes.

1.3 Clipping with Diodes

The grid circuit clipping described in Art. 1.2 is the same in principle as clipping with diodes, since the grid-cathode circuit forms an effective diode. Figure 1.10 shows the circuit diagram of a typical diode clipper. The bias voltage E_{bb} is polarized in such a way that when the signal voltage e_s is zero, the plate is negative with respect to the cathode, and conduction is not possible. The tube remains nonconducting as long as

$$(E_{bb} - e_s) \leq 0$$

However, as soon as the signal voltage e_s exceeds E_{bb}, the plate is positive with respect to the cathode and conduction occurs over that part of the cycle for which this condition exists. The various waveforms appear as shown in Fig. 1.11.

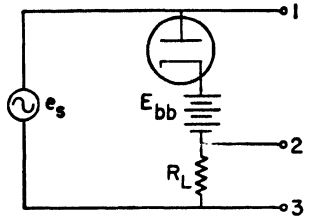


Fig. 1.10. Diode clipper.

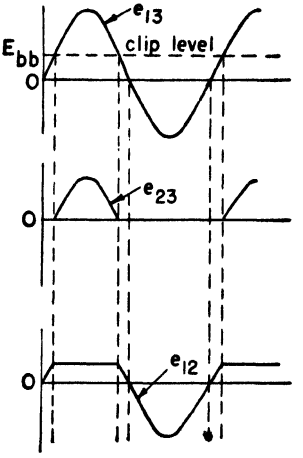


Fig. 1.11. Diode clipper waveforms.

Either half of the input waveform may be clipped depending upon the voltage selected as the output voltage. Changing the magnitude or polarity of E_{bb} merely changes the clip level. Again we note that the tube acts as a switch, being open when nonconducting, and connected through a low resistance r_p when passing current.

1.4 The Square-Wave Generator

At relatively low frequencies, a comparatively simple, but very useful application of the elementary wave-shaping circuits discussed thus far, is the clipper type of square-wave generator. Since the square wave is extremely important in subsequent shaping circuits, its method of generation is worth noting.

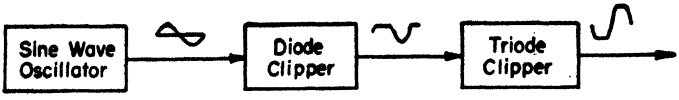


Fig. 1.12. Clipper type of square-wave generator (block diagram).

The functional block diagram of *one* type of square-wave generator is shown in Fig. 1.12. The circuit diagram is shown in Fig. 1.13.

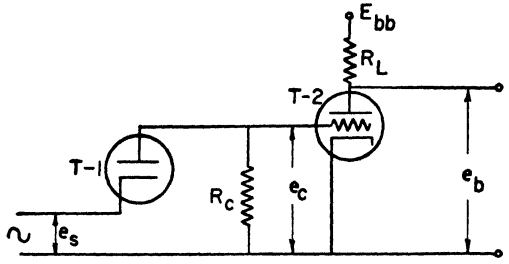


Fig. 1.13. Circuit diagram of clipper type of square-wave generator.

Since the diode conducts only on the negative half cycles of the input signal e_s, the positive half cycles are clipped off. As a result, the grid of the triode is never driven into the positive region, so the grid does not conduct current. Consequently, the grid-to-cathode resistance is substantially an open circuit so that it has no appreciable shunting action on the resistor R_c. As a result, the diode clipper can be isolated for purposes of analysis, as shown in Fig. 1.14. This is the same as the diode clipper discussed in Art. 1.3 except that the bias voltage E_{bb} is zero, causing the clip level to be at 0 volts. Hence, the grid voltage on the triode (T-2) is practically zero at all times except when the diode (T-1) conducts, that is, except when the signal voltage e_s goes negative, as indicated by the waveforms of Fig. 1.15. In other words the diode is operating in the same manner as a half-wave rectifier. The direction of current flow when T-1 conducts is as indicated in Fig. 1.14 so that the output voltage from the diode is a negative half cycle of a sine wave, and the input to the grid of the triode will be a succession of these negative pulses.

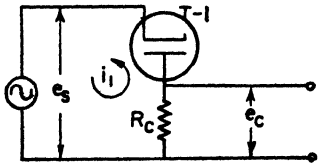


Fig. 1.14. Diode clipper section of square-wave generator.

The operation of the triode clipper may be seen from the typical transfer characteristic shown in Fig. 1.16. Since the grid voltage never goes positive, it is unnecessary to bias the triode and only plate current cutoff clipping is used. Thus, the output voltage from the triode will be squared off, relative to the input, and inverted

with respect to it. To produce a really square wave, a succession of clipping circuits is ordinarily required.

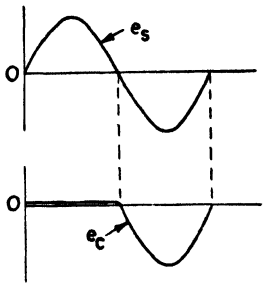


Fig. 1.15. Diode clipper waveforms.

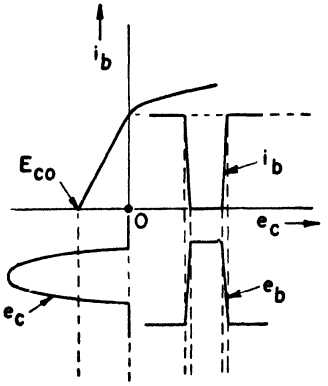


Fig. 1.16. Transfer characteristic for triode clipper and associated waveforms.

1.5 Clipping with Cathode Followers

The cathode follower amplifier finds extensive use in pulse-shaping applications as a matching device between a wave-shaping channel and some low-impedance load, such as the deflecting coils in a magnetically deflected cathode ray tube. It is seldom employed as a wave shaper and clipping is normally undesirable. As a matter of fact, the cathode follower is so difficult to overdrive that it finds a very wide use as a buffer stage, in order to isolate other stages whose operating characteristics would be unfavorably affected by the loading introduced when the grid of a conventional amplifier goes positive. Thus, its superiorities and limitations are worth separate consideration.

The basic mechanism by which clipping does occur in a cathode follower is inherently the same as that of the simple overdriven amplifier, the only real difference lying in the nature of the so-called "input-output characteristic." In a cathode follower of the form

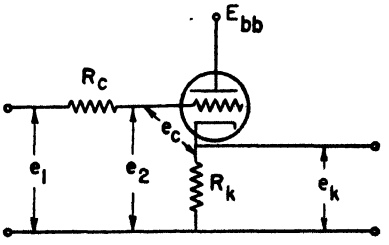


Fig. 1.17. Cathode follower.

shown in Fig. 1.17, this is a relationship between the input voltage e_1 and the output voltage e_k. The corresponding characteristic of the triode amplifier is the transfer characteristic, which gives the relationship between the input (grid voltage) and the output (plate current). The general form of the input-output characteristics for both circuits is substantially the same, the difference being largely in the extent of the linear range of operation.

To facilitate discussion, let us determine a typical input-output characteristic of the cathode follower of Fig. 1.17. Figure 1.18

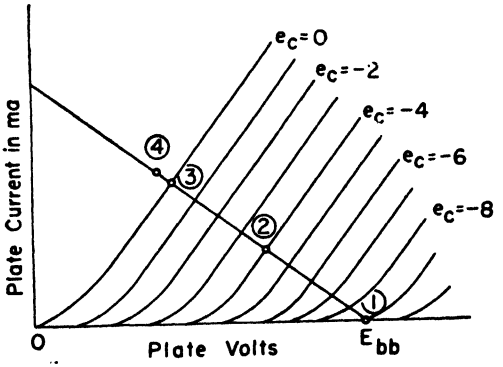


Fig. 1.18. Load line on triode static characteristics for cathode follower, showing four possible operating points.

shows the load line plotted on typical triode static characteristic curves. Since the order of connection of tube and load is irrelevant in graphical solutions of this type, the load line is drawn in the same manner as for a conventional triode amplifier.

The grid voltage (e_c) is the grid-to-ground voltage (e_2) minus the cathode-to-ground voltage (e_k). Hence,

$$e_c = e_2 - e_k \tag{1.2}$$

or, rearranging terms,

$$e_2 = e_c + e_k \tag{1.3}$$

But the drop across the cathode resistance is given by

$$e_k = i_b R_k \tag{1.4}$$

Now, assume several operating points, such as the points labeled (1), (2), and (3) in Fig. 1.18. This gives us values of e_c and i_b, which

allow us to calculate e_k and then e_2. As long as the grid voltage e_c is negative, no grid current flows, there is no voltage drop in the

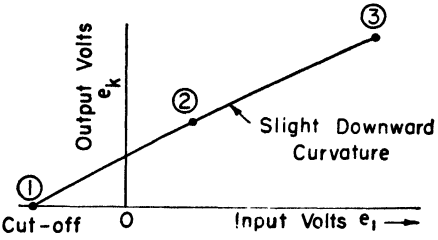


Fig. 1.19. Cathode follower input-output characteristic for linear range of operation.

clipping resistor R_c, and $e_1 = e_2$. Thus the input-output characteristic has the form shown in Fig. 1.19.

Now, when the grid goes positive, as at point (4) in Fig. 1.18, grid current flows producing the typical equivalent grid circuit shown

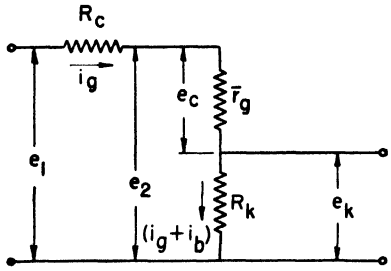


Fig. 1.20. Equivalent grid circuit of cathode follower when overdriven in the positive direction.

in Fig. 1.20. The amount of grid current is given by the expression

$$i_g = \frac{e_c}{\bar{r}_g} \tag{1.5}$$

Since e_c was assumed, and $\bar{r}_g$ is known, then i_g can be calculated. This allows the drop across the clipping resistor to be calculated as

$$e_{R_c} = i_g R_c \tag{1.6}$$

since the input voltage is given by

$$e_1 = e_{R_c} + e_2 \tag{1.7}$$

But

$$e_2 = e_c + e_k = e_c + R_k(i_b + i_g) \tag{1.8}$$

hence, the input voltage is determined according to the relation

$$e_1 = i_g R_c + e_c + R_k(i_b + i_g) \tag{1.9}$$

If several positive values for the grid voltage e_c are assumed, the complete input-output characteristic will appear as shown by Fig. 1.21. The value of e_1 corresponding to point (3) on the curve is

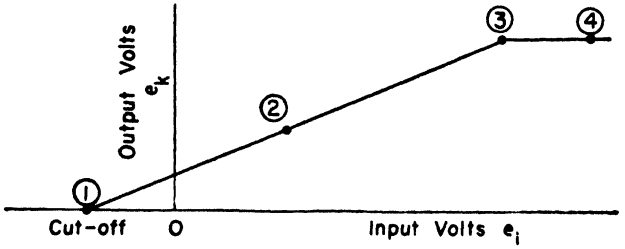


Fig. 1.21. Complete input-output characteristic for the cathode follower of Fig. 1.17.

generally of the order of 250 volts. Thus, it is apparent that the characteristic is very linear over an extremely wide range of values of input voltage, whereas grid clipping in a conventional triode amplifier starts virtually as soon as the *grid* goes positive. Consequently, it is exceedingly difficult to overdrive a cathode follower and the reason for its desirability as a coupling device or buffer is apparent.

Typical waveforms are shown in Fig. 1.22. Note that e_2 follows the input voltage e_1 except during grid current flow, but that e_k, the output voltage, follows the input only over the linear region of

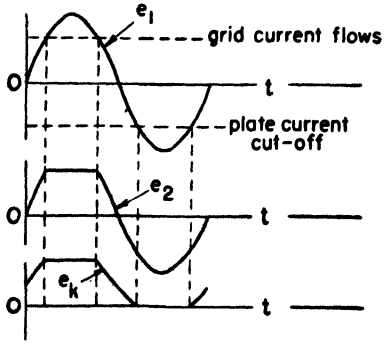


Fig. 1.22. Waveforms when cathode follower is overdriven.

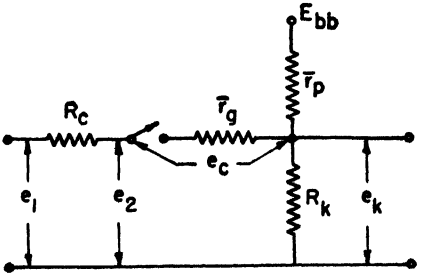


Fig. 1.23. Complete equivalent circuit of cathode follower, for overdriven operation.

the input-output characteristic. The complete equivalent circuit of the cathode follower is as indicated in Fig. 1.23.

1.6 Differentiators and Integrators (Transient Analysis)

The complexity of the waveforms required in most cathode ray tube applications is such that circuits composed of tubes and resistors only cannot produce the necessary shaping operations. The peculiar transient responses of the simple series *R-C*, *R-L*, and *R-L-C* circuits are then employed and the tube merely fulfills the function of a switch. In most cases, the required waveforms must be repeated periodically in time so that the switch tube must be operated at the desired repetition rate. Generally, a square-wave generator is used to establish this switching frequency by applying it to the grid of the switch tube.

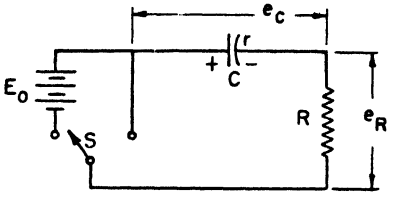


Fig. 1.24. Series *R-C* circuit.

It is left as an exercise for the student to show that the following expressions are true and indicate the transient response of the circuit shown in Fig. 1.24.

$$e_R = (E_0 - r)\epsilon^{-t/RC} \tag{1.10}$$

$$e_c = E_0 - (E_0 - r)\epsilon^{-t/RC} \tag{1.11}$$

where E_0 is the applied voltage during the interval under consideration and r, the initial condenser voltage. If the switch S is operated periodically at a frequency f, the waveform of the voltage applied to the circuit is as shown in Fig. 1.25.

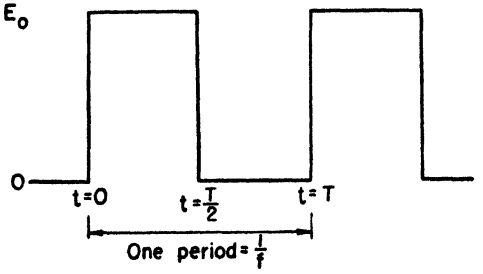


Fig. 1.25. Square-wave voltage applied to series *R-C* circuit.

There are three distinguishable cases depending upon the relationship between the half period of the square wave ($T/2$) and the time constant (RC) of the circuit, as follows:

(1) $T/2 \gg RC$ (i.e., $T/2$ is at least 5 time constants).
(2) $T/2 \approx RC$.
(3) $T/2 \ll RC$ (i.e., $T/2$ is not more than 1/5 time constant).

The waveforms obtained for each of these conditions are shown in Fig. 1.26 for the steady-state condition.

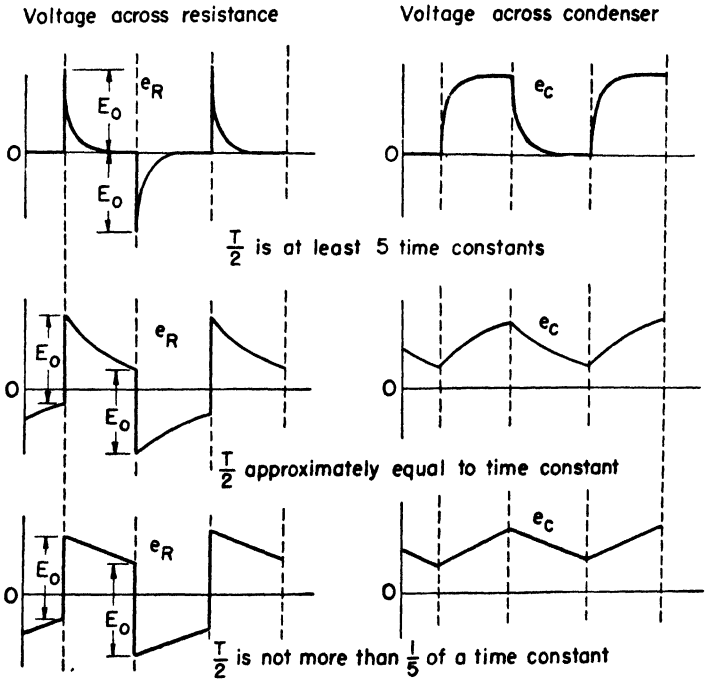


Fig. 1.26. Steady-state waveform in series R-C circuit when a periodic square wave of period T is applied.

At this point it is worthwhile to inquire about the nature of the waveforms if the input voltage had been symmetric about the zero voltage axis. The correct waveforms are shown in Fig. 1.27. This point is raised because the impression frequently develops that the voltage peaks of the voltage across the resistor (e_R) are only half the height of those in Fig. 1.26, due to the symmetric nature of the input

voltage. This erroneous conclusion arises from superficial consideration of the problem, since the transient response is governed by the magnitude of the *change* in voltage. In both cases the change in

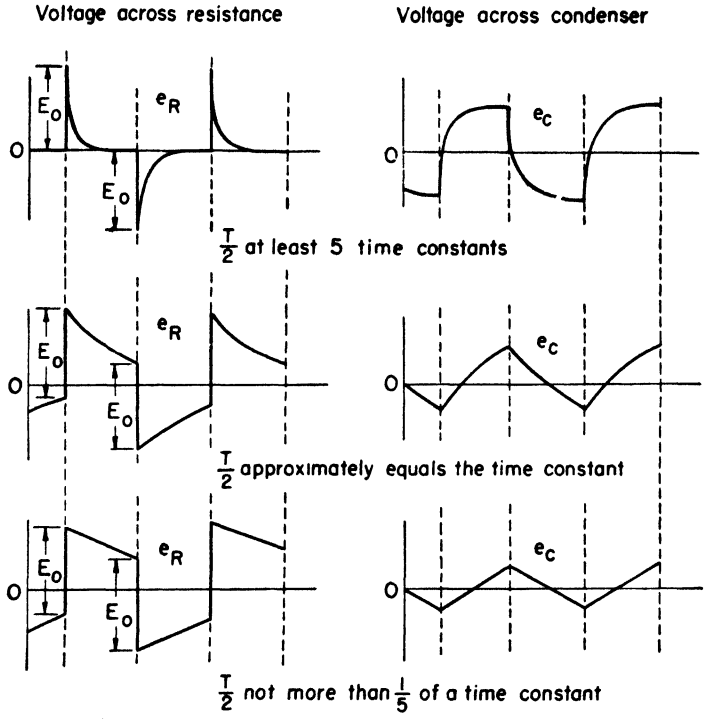


Fig. 1.27. Steady-state waveforms in a series *R-C* circuit when a balanced, symmetrical square wave is applied.

voltage is the same and the waveform of the transient current or voltage across the resistance must be the same. The difference exists in the condenser voltage waveform as one would expect.

There are several important observations connected with these waveforms. First, if an *R-C* circuit were to be used as a coupling device between two other networks, and minimum distortion of the input square wave were a requirement, then a long time constant (*RC*), compared to the half period of the square wave, would be necessary if the output voltage were taken across the resistor. On the other hand, if the output were taken across the capacitance, a

short time constant would be necessary for transmission with the same degree of fidelity.

Another point worth noting is that when the time constant is small compared to the square-wave half period, the voltage across the resistor is *peaked* compared to the input. Considering the circuit differential equation for the series R-C circuit of Fig. 1:24,

$$E_0 = \frac{1}{C}\int i\,dt + Ri \tag{1.12}$$

Multiplying through by C yields

$$CE_0 = \int_0^t i\,dt + RCi \tag{1.13}$$

But, since the time constant (RC) is very small compared to the half period ($T/2$) for the peaker circuit, then

$$\int_0^t i\,dt \gg (RC)i \tag{1.14}$$

Consequently, as an approximation,

$$CE_0 = \int_0^t i\,dt \tag{1.15}$$

Take the derivative of both sides of this equation with respect to time, obtaining

$$C\frac{dE_0}{dt} = i \tag{1.16}$$

but

$$i = \frac{e_R}{R} \tag{1.17}$$

Rearranging terms yields

$$e_R = RC\frac{dE_0}{dt} \tag{1.18}$$

This last equation indicates that the voltage across the resistance is proportional to the derivative of the input under the assumed conditions. Thus, Fig. 1.28 is the circuit diagram of an R-C differentiator circuit. Figure 1.29 shows photographs made of the output from a circuit of this type.

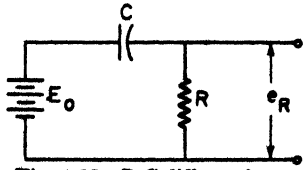


Fig. 1.28. *R-C* differentiator.

Now, consider the reverse situation, that is, the voltage across the capacitance when the time constant is very large compared to

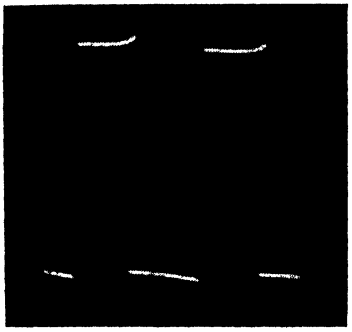


Fig. 1.29(a). Square wave applied to *R-C* differentiator.

Fig. 1.29(b). Output from *R-C* differentiator when a square wave is applied. Note the double deviation from the ideal.

Fig. 1.29(c). Same as Figure 1.29(b) except that the time constant has been reduced. Note how much closer it appoximates the ideal differentiating action.

the square-wave half period. Then, by the same method as for the differentiator, we obtain

$$CE_0 = \int_0^t i\,dt + (RC)i \tag{1.19}$$

But, since the time constant is very large compared to the half period of the square wave, then

$$(RC)i \gg \int_0^t i\,dt \tag{1.20}$$

As a result, the term containing the integral can be neglected in comparison to the other terms, leaving

$$CE_0 = (RC)i \tag{1.21}$$

Integrate both sides of this equation with respect to time, obtaining

$$C\int_0^t E_0\,dt = RC\int_0^t i\,dt \tag{1.22}$$

but, the current i is given by

$$i = C\frac{de_c}{dt} \tag{1.23}$$

or, integrating this,

$$\int_0^t i\,dt = Ce_c \tag{1.24}$$

which, when substituted into Eq. (1.22), yields

$$C\int_0^t E_0\,dt = RC^2e_c \tag{1.25}$$

or, rearranging terms,

$$e_c = \frac{1}{RC}\int_0^t E_0\,dt \tag{1.26}$$

Consequently, it is apparent that the voltage across the capacitor is the integral of the input, under the assumed conditions, and Fig. 1.30 is the circuit diagram of an *R-C integrator*.

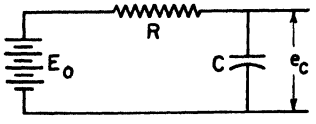


Fig. 1.30. *R-C* integrator.

An exactly similar analysis, when applied to the series R-L circuit, produces the results illustrated in Fig. 1.31.

The approximations used in obtaining these results should be carefully noted. Although the approximations generally hold, con-

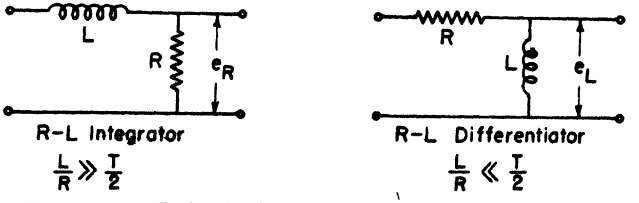


Fig. 1.31. *R-L* circuits for integrating and differentiating.

ditions arise where they are not fulfilled and this possibility cannot be disregarded.

1.7 Differentiators and Integrators (Fourier Analysis)

The preceding method of analysis of simple networks is usually the quickest way to a direct solution of the problem at hand. However, it may not always yield information in the form sought by the engineer. In a great many cases the attenuation and phase shift of the circuit as a function of frequency are bits of data very necessary to the design of allied equipment, and the Fourier series method of

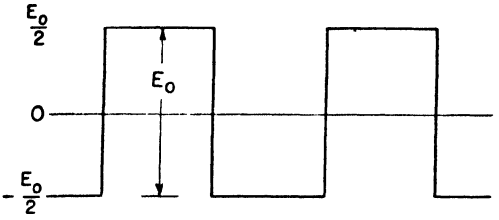


Fig. 1.32. Balanced symmetrical square wave applied to an *R-C* integrator.

analysis is more convenient. The procedure then entails three distinct steps:

(1) Fourier series analysis of the input signal.
(2) Analysis of the circuit in terms of conventional a-c theory.
(3) Synthesis of the harmonics in the output into a Fourier series.

As an example, consider the case of a square wave, of the form shown in Fig. 1.32, applied to the *R-C* integrating circuit of Fig. 1.33. The Fourier analysis of the square-wave input is given by Eq. (1.27) below.

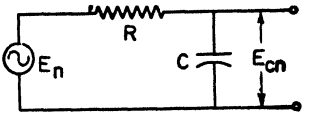


Fig. 1.33. *R-C* integrating circuit.

$$e_i = \frac{4E_0}{\pi}\left(\sin \omega t + \frac{1}{3}\sin 3\omega t + \frac{1}{5}\sin 5\,\omega t + \cdots\right) \quad (1.27)$$

where E_0 is the peak-to-peak amplitude of the wave; ω, the frequency of the fundamental component; and n, the order of the harmonic (1, 2, 3, $\cdots$). Then for any given harmonic of order n

$$E_n = IZ = I\left(R - \frac{j}{n\omega C}\right) \quad (1.28)$$

The output voltage, which is the voltage across the condenser, is given by

$$E_{cn} = IX_c = -I\left(\frac{j}{n\omega C}\right) \tag{1.29}$$

Hence, the complex voltage ratio of output to input is

$$\frac{E_{cn}}{E_n} = -\frac{j/n\omega C}{R - (j/n\omega C)} = -\frac{j}{n\omega RC - j} \tag{1.30}$$

Divide through by $(-j)$ to obtain

$$\frac{E_{cn}}{E_n} = \frac{1}{1 + jn\omega RC} \tag{1.31}$$

This ratio may also be written in polar form as

$$\frac{E_{cn}}{E_n} = A_n = |A_n| \underline{|-\beta_n} = \epsilon^{-(\alpha+j\beta)} \tag{1.32}$$

where α is the attenuation in nepers and β, the phase lag in radians. Hence

$$A_n = \frac{1}{1 + jn\omega RC} = \frac{1}{\sqrt{1 + (n\omega RC)^2}\,\underline{|\beta_n}} \tag{1.33}$$

where $$B_n = \tan^{-1} n\omega RC \tag{1.34}$$

so $$A_n = \left(\frac{1}{\sqrt{1 + (n\omega RC)^2}}\right)\underline{|\ tan^{-1}\ n\omega RC} \tag{1.35}$$

Therefore, the absolute value of the complex voltage ratio is

$$|A_n| = \frac{1}{\sqrt{1 + (n\omega RC)^2}} = \frac{1}{\sqrt{1 + \left(\frac{n\omega}{\omega_0}\right)^2}} \tag{1.36}$$

and $$B_n = tan^{-1}\left(\frac{n\omega}{\omega_0}\right) \quad \text{where} \quad \omega_0 = \frac{1}{RC} \tag{1.37}$$

The frequency characteristics of the network are as shown in Fig. 1.34. The significance of the quantity ω_0 is that at this frequency, the amplitude is down to 0.707 of its maximum value. In other words, it is the half-power frequency. This type of relationship will have particular significance in the material to be presented subsequently on video amplifiers.

Hence, we can conclude that the fundamental will come out of the network with its amplitude multiplied by a fractional quantity A_1

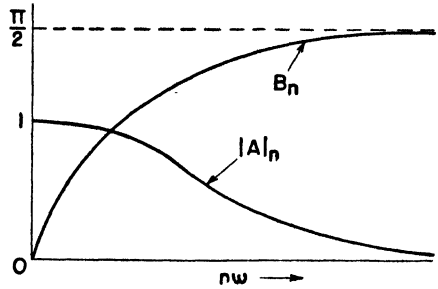


Fig. 1.34. Network characteristics as a function of frequency for the circuit of Fig. 1.33.

and its phase angle retarded by B_1, and so on for each of the harmonics. The sum of all of these modified terms will be the output voltage E_c.

In the case of an integrator, the time constant RC is very large compared to the square-wave half period $T/2$. Hence,

$$\omega \gg \omega_0$$

so
$$|A|_n = \frac{\omega_0}{n\omega}$$

and
$$B_n = tan^{-1}\, n(\infty) = \pi/2$$

Therefore, the amplitude of each harmonic coming out of the network is reduced by an amount proportional to the order of the harmonic. Furthermore, each successive harmonic has its phase angle retarded by $\pi/2$, thus changing the original sine series to a negative cosine series so that

$$e_c = -\left(\frac{4E_0}{\pi}\right)\left(\frac{\omega_0}{\omega}\right)\left(\cos \omega t + \frac{1}{9}\cos 3\,\omega t + \frac{1}{25}\cos 5\omega t + \cdots\right) \quad (1.38)$$

This is recognizable as the Fourier development of a symmetrical (back to back) saw tooth, which is the same as the waveform of Fig. 1.26.

This method of analysis, while admittedly somewhat longer, yields information in an exceptionally convenient form. This advantage is somewhat nullified in many cases, however, since the re-

sulting waveform obtained by conventional transient analysis, as given in Art. 1.6, could be broken up into its Fourier components with comparatively little difficulty. It is important, nevertheless, to be fully cognizant of both methods of attack and to understand their interrelationships.

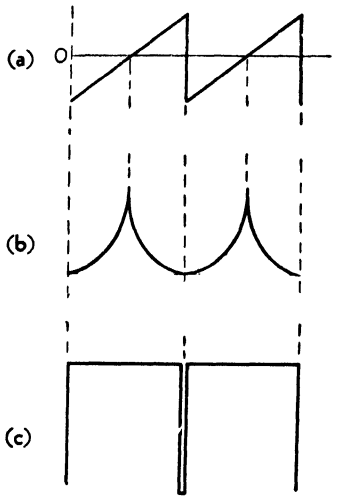


Fig. 1.35. (a) Saw-tooth voltage applied. (b) Integrated output. (c) Differentiated output.

1.8 Response of Differentiators and Integrators to Unusual Waveforms

So far, the discussion of differentiators and integrators has been limited to their responses to square-wave inputs. Recognizing their abilities to differentiate or integrate, their behavior in response to any waveform may be determined. A typical example will be discussed as an illustration.

Assume that a saw-tooth voltage of the form shown in Fig. 1.35(a) is applied to the usual *R-C* integrator. The expression for the input voltage may be written as

$$e_i = Kt \qquad \text{where } K = \text{constant} = \text{slope}$$

Consequently, the output voltage is given by

$$e_0 = \frac{1}{RC}\int e_i\,dt = \frac{1}{RC}\int (Kt)\,dt \tag{1.39}$$

or

$$e_0 = \frac{1}{RC}\left(\frac{Kt^2}{2} + K'\right) \tag{1.40}$$

This is the equation for a parabola and the waveform appears as shown in Fig. 1.35(b). Similarly, for the differentiator,

$$e_i = Kt \tag{1.41}$$

so

$$e_0 = RC\,\frac{de_i}{dt} = (RCK) \tag{1.42}$$

which is a constant, and the waveform appears as shown in Fig. 1.35(c).

This same method may be applied to any input waveform. Of course, the results are only approximations, subject to the limitations imposed as discussed in Art. 1.6, but in most cases, sufficient accuracy results to warrant the approximation.

1.9 Removal of the D-C Component

A characteristic of the simple *R-C* circuits discussed in Art. 1.6, which consistently causes difficulty in the design of pulsed circuits, is the removal of the d-c component of the input signal. The effect is of considerable importance in circuits where the bias required for proper operation is critical. For example, consider a recurrent pulse of the form shown in Fig. 1.36, applied to an *R-C* coupling circuit.

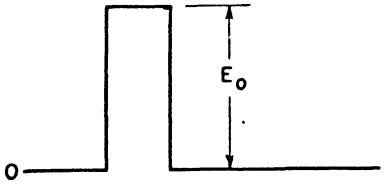


Fig. 1.36. Recurrent pulse applied to an *R-C* coupling circuit.

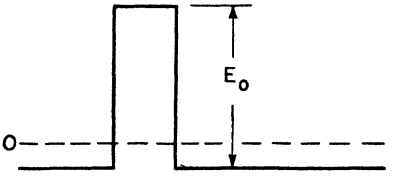


Fig. 1.37. Output of coupling circuit when the recurrent pulse of Fig. 1.36 is applied.

The waveform at the output of the coupling circuit will appear as shown in Fig. 1.37, where the waveform has averaged itself about the zero axis in such a way that the areas above and below the zero voltage axis are equal. Thus, the d-c component of the input has been removed.

A Fourier analysis of the input and output waveforms will be identical except that the input will contain a constant term which will be absent in the output. The application of a voltage pulse to the *R-C* circuit generates a transient so that the output waveform appears as shown in Fig. 1.38 during the transient period. Ap-

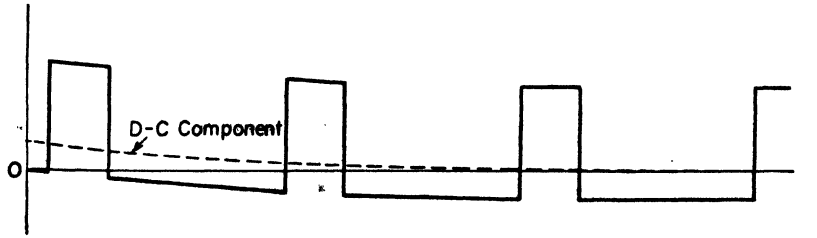


Fig. 1.38. Output from coupling circuit during the transient period.

parently then, the steady-state recurrent pulse is superimposed upon the conventional exponential transient. Thus, the shifting of the waveform so that its average value is zero is accomplished by the transient current which flows to charge the coupling condenser.

If the pulse widths are changed, the average d-c value at the input would be changed, and the output waveform would not be oriented exactly as it was before. This is illustrated in Fig. 1.39.

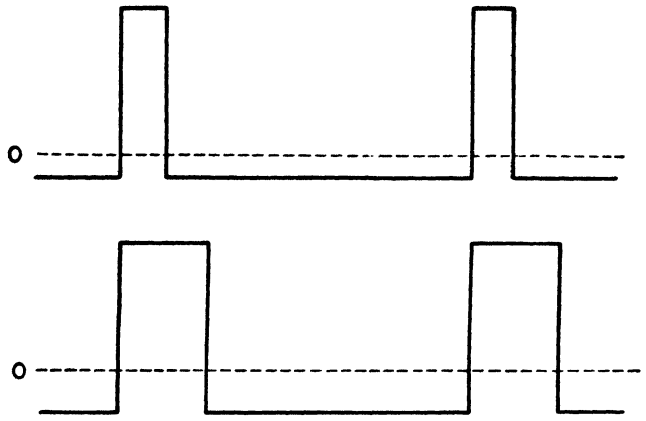


Fig. 1.39. Effect on coupling circuit output due to changes in pulse width.

In general, in pulse circuits, measurements are made with respect to the base of the waveform, not with respect to the comparatively inaccessible average value. However, all biases must be calculated with respect to the average value and it is here that difficulty most frequently arises.

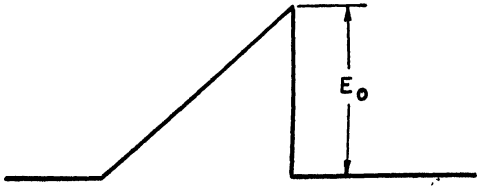


Fig. 1.40. Saw-tooth voltage.

For example, consider the saw tooth of Fig. 1.40, and assume that it is applied through an *R-C* coupling circuit to a triode employing

grid clipping. The waveform is to be clipped in such a way that the voltage directly on the grid of the tube is to appear as shown in Fig. 1.41. The flat-topped part δ is assumed to be a very critical time

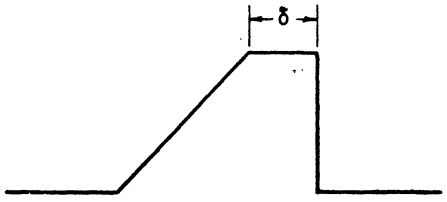


Fig. 1.41. Voltage required directly on the grid of a triode clipper.

from which a small deviation cannot be tolerated. The first tendency is to consider applying sufficient bias (E_{cc}) to the clipper so that the waveform on the grid appears on the transfer characteristic as shown in Fig. 1.42. This would result in a large error in the length of δ because the bias was not computed with respect to the average d-c value. The proper way to apply the bias is shown by Fig. 1.43.

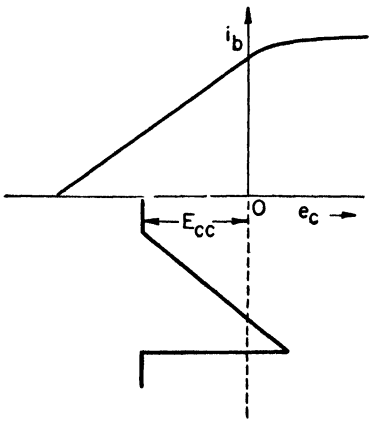


Fig. 1.42. Incorrect method of biasing as it appears on triode transfer characteristic.

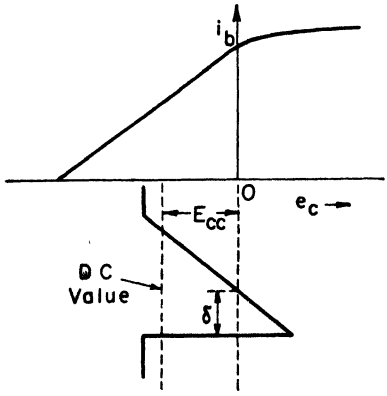


Fig. 1.43. Proper method of calculating bias, as it appears on the triode transfer characteristic.

1.10 D-C Restoration and the Clamper

The removal of the d-c component, as discussed in Art. 1.9 is, as previously pointed out, an important consideration in a great many design applications where the d-c component of the input waveform

is necessary for proper operation of a succeeding component in the wave-shaping channel. Thus, it frequently becomes necessary to restore the original d-c component, or some other d-c voltage. The wave-shaping circuits which perform this function are called *d-c restorers* or, quite commonly, *clampers*.

There may be some doubt as to the actual necessity of producing a separate circuit to restore the d-c component when the insertion of a constant biasing voltage would, in effect, provide whatever average value was required. Such a scheme would perform effectively only so long as the incoming signal remained unchanged in shape and magnitude, since variations in either would produce changes in the average value necessitating a change in the requirements placed upon the biasing system for each different input waveform. As a consequence, even a very elaborate bias circuit would experience difficulty in properly compensating for waveform variations. It would be much more desirable to add in the d-c component in a manner which would be substantially unaffected by variations in the input signal. This is the function of the *clamper*.

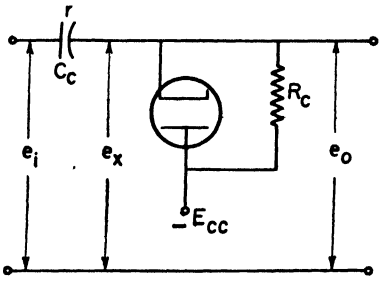


Fig. 1.44. Diode clamper.

There are several types of d-c restorers, but simplest in operation and in concept is the diode clamper shown in Fig. 1.44. Since the diode has a conducting path around it, through R_c, that side of the coupling condenser C_c is nominally held at $-E_{cc}$ volts. Hence, the voltage across the coupling condenser is

$$e_i(\text{min}) + E_{cc}$$

The circuit is to *clamp* the base of the output voltage of the clamping circuit to the voltage $-E_{cc}$. When the input voltage jumps to e_i (max), over a total change in voltage of E_m, the output voltage e_0 does the same thing. Refer to Fig. 1.45. The diode does not conduct because the plate is still negative with respect to the cathode. The time constant R_cC_c is made very long, by making R_c large, so that no appreciable change in voltage occurs across the condenser. Thus, when the input voltage drops back to e_i (min), the output drops back to $-E_{cc}$. No matter how the waveform varies, as long

as the input voltage does not drive the cathode voltage below the plate voltage ($-E_{cc}$), the output waveform is clamped at $-E_{cc}$, just as the input is clamped at e_i(min) by the source.

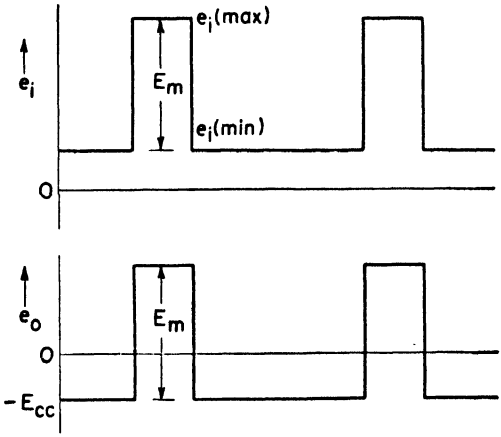


Fig. 1.45. Diode clamper waveforms.

However, suppose the signal does drive the cathode negative with respect to the plate, by dropping below e_i(min). Then, the diode conducts, becoming a very low resistance $\bar{r}_p$, compared to R_c, and the condenser very quickly charges back to a value such that, except during this charging period, the output is kept clamped at $-E_{cc}$. Thus, in general, the output is clamped at the reference voltage regardless of the changes in waveform or dimensions of the input. However, when the duration of the signal being clamped is of the order of magnitude of the charging time constant ($\bar{r}_p C_c$), some undesirable distortion in the waveform may occur.

Although triodes are often used as clampers, the fundamental principle of operation is the same, so that the discussion of this article is of sufficient generality to cover most of the circuits usually encountered.

1.11 Grid Leak Bias

The preceding two articles have dealt with the removal and restoration of d-c voltages after passing through R-C coupling circuits. When waveforms of the type discussed in the preceding articles are applied directly to the grid of an amplifier, another effect, called

grid leak bias, occurs. It consists of an actual average d-c voltage across the resistance in the coupling circuit.

Assume that the input to a typical triode amplifier is a symmetrical square wave as shown in Fig. 1.46. Note that there is no d-c

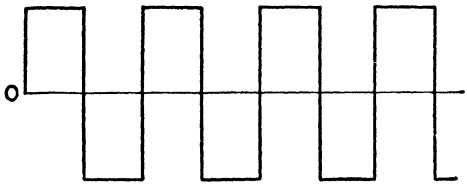


Fig. 1.46. Square wave applied.

component. During the first positive half cycle, the grid goes positive with respect to the cathode and the coupling condenser tends to charge through the static grid resistance $\bar{r}_g$. The equivalent grid charging circuit is shown in Fig. 1.47. Hence, the charging time constant is

$$T_{ch} = C_c\left(\frac{\bar{r}_g R_g}{\bar{r}_g + R_g}\right) \tag{1.43}$$

which is quite small since $\bar{r}_g$ is small compared to R_g. On the negative

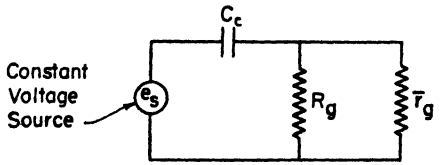


Fig. 1.47. Equivalent grid circuit when grid voltage is positive.

half cycle of the input, no grid current flows and the coupling condenser tends to discharge through the grid leak resistor R_g with a time constant

$$T_{dsch} = C_c R_g \tag{1.44}$$

which is comparatively large. Due to this difference in time constants during the transient period, more charge collects on the condenser during charge than leaks off during discharge. As a result, the voltage across the resistor R_g developes a mean d-c bias as shown in Fig. 1.48, such that in the steady state, the positive

excursion of the grid voltage is just sufficient to overcome the losses in the grid circuit during discharge.

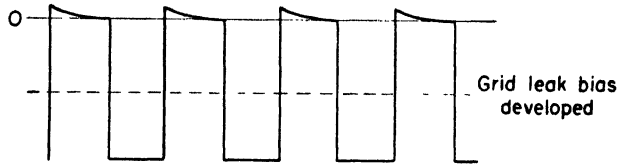


Fig. 1.48. Grid leak bias.

1.12 Analysis of a Typical Pulse Generator

At this stage it would be advisable to recapitulate and discuss a wave-shaping system involving a consideration of most of the principles outlined so far. Such a procedure, speaking figuratively, will provide a glimpse of the forest as well as the trees, and will certainly facilitate subsequent discussions which involve the same general method of attack.

Let it be required to produce a series of positive pulses with a given spacing T_r, and clamped at a voltage $-E_{cc}$. The only readily available signal source is a sine wave oscillator.

The first step in design is to break the system down into its fundamental components in the form of a functional block diagram.

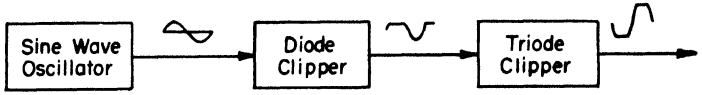


Fig. 1.49. Block diagram of clipper type of square-wave generator.

In a great many applications it is highly desirable to produce a square wave from the original sine wave, since this is an excellent starting point for further shaping operations. As a matter of convenience, let us use the square-wave generator discussed in Art. 1.4 and whose block diagram is reproduced in Fig. 1.49. Since the output pulses are to be spaced at intervals of T_r, the frequency of the sine wave oscillator should be set to $f_r = 1/T_r$. The requirement that the final output is to consist of pulses means that a narrowing or peaking action must be introduced as shown in Fig. 1.50. In addition, the output pulses are required to be of a single polarity so that clipping is again necessitated in order to remove the pulses of the undesired

polarity. The two final requirements are that the pulses be positive and be clamped at $-E_{cc}$. If grid circuit clipping is used in the last

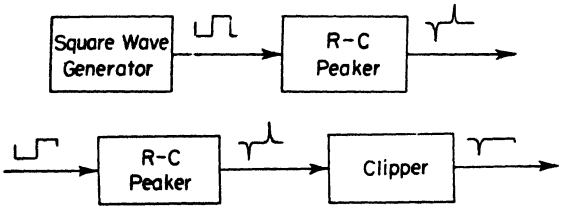


Fig. 1.50.

clipping stage, the output pulses *will* be positive due to phase inversion through the triode. Thus, the complete system block diagram appears as shown in Fig. 1.51. This diagram does not, by

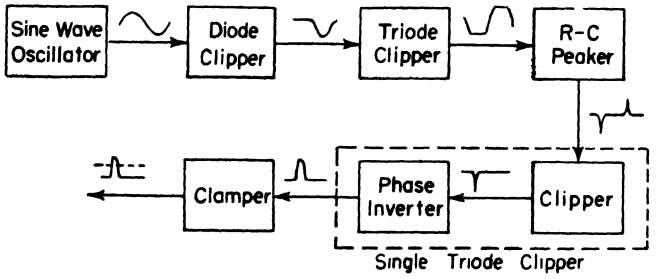


Fig. 1.51. Complete system block diagram of pulse generator.

any means, represent the ideal solution, since there are a great many ways of obtaining this same result.

The analysis of the square-wave generator will be omitted, since

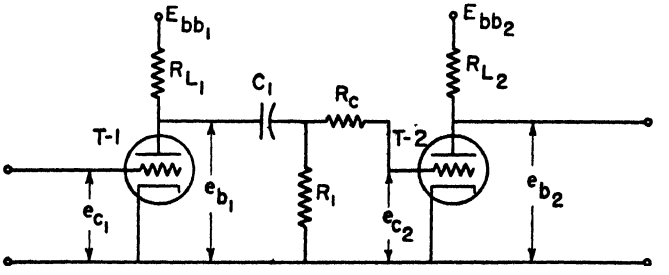


Fig. 1.52. Circuit diagram of *R-C* peaker and clipper-phase inverter, including the last tube in the square-wave generator.

it was covered in detail in Art. 1.4. However, consider the next three components, the *R-C* peaker, and the clipper-phase inverter. The circuit diagram including the last tube of the square-wave generator is shown in Fig. 1.52. Referring to the waveform of Fig. 1.53, which is a rather idealized representation of the plate voltage

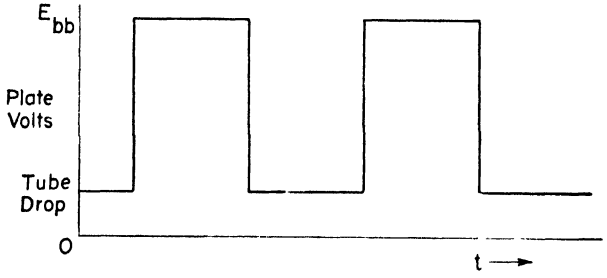


Fig. 1.53. Idealized output from square-wave generator.

from the square-wave generator, we see that the application of this square wave to the differentiating circuit will result in the formation of positive peaks of sufficient amplitude to drive the grid of the second tube (T-2) positive, producing grid circuit clipping and introducing a low shunting resistance $\bar{r}_{g_2}$ across R_1 and R_c during such intervals. Thus the equivalent plate circuit of T-1 and equivalent grid circuit of T-2 appear as shown in Fig. 1.54. Since the two

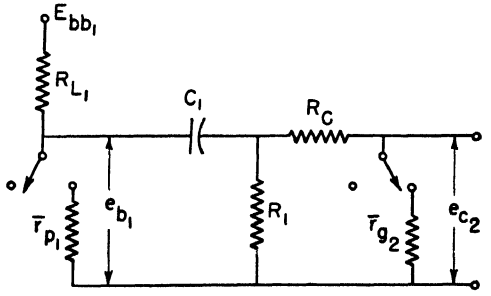


Fig. 1.54. Equivalent plate circuit of tube *T*-1.

switches operate in synchronism, there are two possible reduced equivalent circuits, one giving the charging circuit for C_1, and the other for the discharging circuit. This circuit breakdown is shown in Fig. 1.55. Applying Thevenin's theorem to the discharging circuit

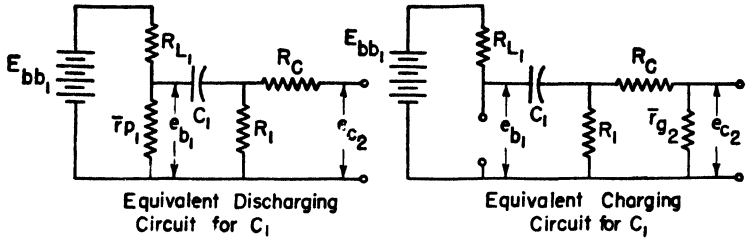


Fig. 1.55. Equivalent plate circuits for the charging and discharging of condenser C_1.

and simplifying the charging circuit by considering the static grid resistance to be negligible compared to the grid-clipping resistor, yields the circuits of Fig. 1.56, where

$$R_e = \left(\frac{\bar{r}_{P_1} R_{L_1}}{\bar{r}_{P_1} + R_{L_1}}\right) \qquad R_t = R_e + R_1 \qquad E_e = E_{bb_1}\left(\frac{\bar{r}_{P_1}}{\bar{r}_{P_1} + R_{L_1}}\right)$$

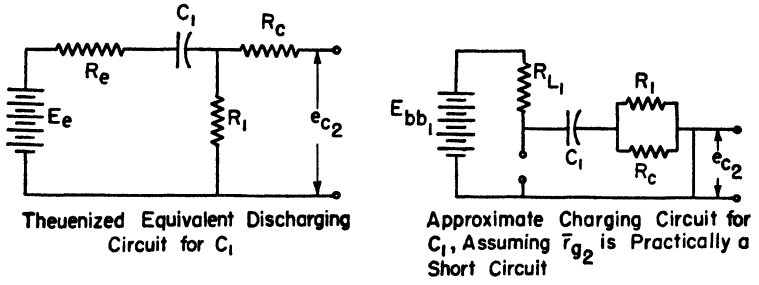


Fig. 1.56. Reduced equivalent circuits.

Thus, when T-1 is made conducting by its grid voltage exceeding cutoff, analysis is made using the equivalent discharging circuit. Assuming the charging time constant to be small compared to the square-wave half period, then we can consider the charging transient to be substantially complete when T-1 is cut on. Consequently, the initial condenser voltage for the discharging transient must be the plate supply voltage E_{bb_1} The equation for the discharging transient current is then

$$i_{dsch} = \left(\frac{E_e - E_{bb_1}}{R_T}\right)\epsilon^{-t/R_T C_1} \tag{1.45}$$

Hence, the grid voltage on T-2 is

$$e_{c2} = i_{\text{dsch}}R_1 = (E_e - E_{bb_1})\left(\frac{R_1}{R_T}\right)\epsilon^{-t/R_TC_1} \tag{1.46}$$

and the voltage across the condenser in the peaking circuit is

$$e_{cx} = E_e - (E_e - E_{bb_1})\epsilon^{-t/R_TC_1} \tag{1.47}$$

Consequently, the plate voltage on T-1 is given by

$$e_{b_1} = e_{cx} + e_{c2} \tag{1.48}$$

and substituting for e_{cx} and e_{c2} yields

$$e_{b_1} = E_e - (E_e - E_{bb_1})\left(1 - \frac{R_1}{R_T}\right)\epsilon^{-t/R_TC_1} \tag{1.49}$$

where $E_e = E_{bb_1}\left(\dfrac{\bar{r}_{P_1}}{\bar{r}_{P_1} + R_{L_1}}\right)$ and $R_t = R_1 + \left(\dfrac{\bar{r}_{P_1}R_{L_1}}{\bar{r}_{P_1} + R_{L_1}}\right)$ (1.50)

Hence, when $t = 0$ at the beginning of the discharging transient, the grid voltage on T-2 is

$$e_{c2} = -\Delta E = (E_e - E_{bb_1})\frac{R_1}{R_T} \tag{1.51}$$

Similarly, the plate voltage on T-1 is

$$e_{b_1} = E_{bb_1} - \Delta E \tag{1.52}$$

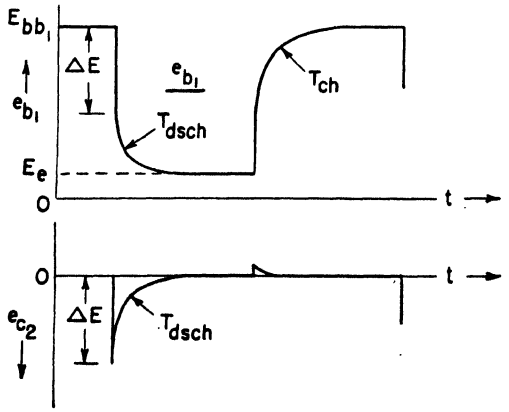


Fig. 1.57. Pulse generator waveforms.

When the transient is complete

$$e_{c_2} = 0 \tag{1.53}$$

and the plate voltage on T-1 is

$$e_{b_1} = E_e \tag{1.54}$$

The time constant of the intervening exponential is

$$T_{\text{dsch}} = R_t C_1 \tag{1.55}$$

The waveforms for this half cycle are plotted and labelled in Fig. 1.57.

When T-1 cuts off, the charging equivalent circuit applies. The initial condenser voltage for the charging half cycle is E_c, the drop across the tube. Following the same method used to calculate the critical points on the preceding half cycle of the waveforms, the plate and grid voltages may be shown to be given by the following expressions:

$$e_{b_1} = E_{bb_1} - (E_{bb_1} - E_e)\epsilon^{-t/R_T C_1} \tag{1.56}$$

where $R_{L_T} = R_{L_1} + \left(\dfrac{R_1 R_c}{R_1 + R_c}\right)$ and $e_{c_2} = 0$

since $\bar{r}_{g_2} \ll R_c$

The initial plate voltage is, of course,

$$e_{b_1} = E_e \tag{1.57}$$

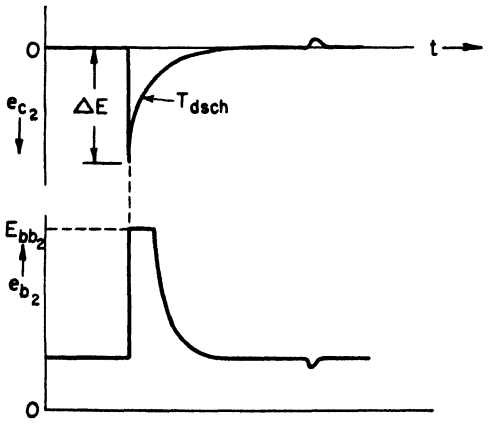


Fig. 1.58. Pulse generator waveforms.

Whereas, the final voltage is

$$e_{b_1} = E_{bb_1} \tag{1.58}$$

The time constant of the intervening transient is

$$T_{ch} = R_{LT}C_1 \tag{1.59}$$

The complete waveforms are given in Fig. 1.58.

The application of this signal voltage to the grid of T-2 will produce the plate voltage waveform e_{b_2} shown in Fig. 1.58. This waveform may be obtained directly from the tube characteristics, but is most readily obtained by consideration of the tube transfer characteristic as shown in Fig. 1.59. Application of this waveform to the diode clamper of Art. 1.10 will produce the required signal. The slight deviation of e_{c_2} from the calculated waveform arises from the assumption that $\bar{r}_{g_2}$ is substantially a short circuit compared to R_c.

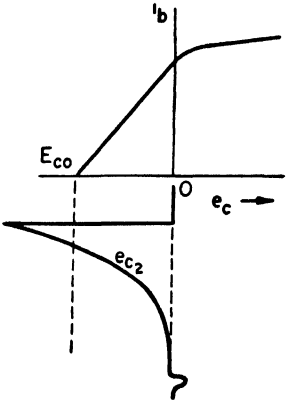


Fig. 1.59. Transfer characteristic for triode clipper in the pulse generator.

1.13 The Saw-tooth Voltage Generator

Most of the material immediately preceding this article has dealt with differentiators and coupling circuits, whereas, for the next few articles, we shall be concerned with the integrating action of the *R-C* circuit. From Art. 1.6 it will be recalled that when a square wave was applied to an *R-C* integrating circuit, the output voltage had the waveform shown in Fig. 1.60. Due to the half-wave symmetry about a line such as t_1, this is usually referred to as a "back-to-

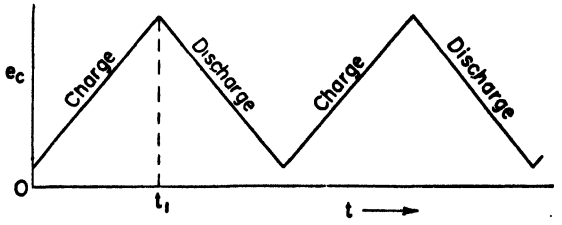


Fig. 1.60. Symmetrical back-to-back saw-tooth voltage.

back saw-tooth" voltage. The symmetry results from equal charging and discharging time constants for the condenser.

A more widely used waveform is the simple saw tooth of Fig. 1.61 which finds extensive application as a sweep voltage in electrostatically deflected cathode ray tubes. The production of such waveforms hinges upon having a very long charging time constant, and

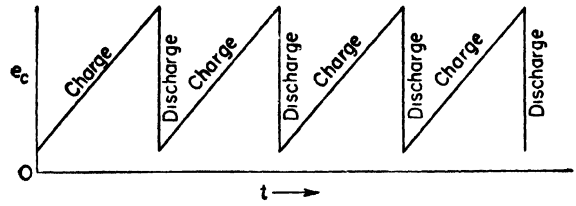


Fig. 1.61. Simple, idealized saw-tooth voltage.

an extremely short discharging time constant. Such a circuit could be obtained as shown in Fig. 1.62 where the switch shorts the condenser, providing a short time constant discharging path. When the switch is open, the condenser charges through the resistor R, which provides the long time constant charging path. The switch is operated periodically to obtain the desired saw-tooth frequency.

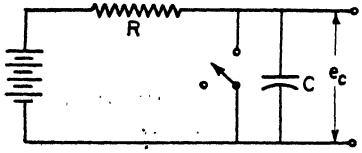


Fig. 1.62. Prototype saw-tooth voltage generator.

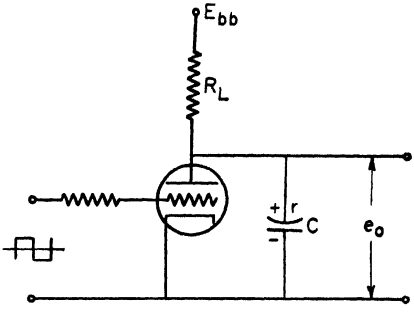


Fig. 1.63. Circuit diagram of saw-tooth generator.

This circuit could be closely approximated by replacing the switch with a vacuum tube which is being overdriven in both directions. The circuit diagram then appears as shown in Fig. 1.63. The equivalent circuit is given in Fig. 1.64.

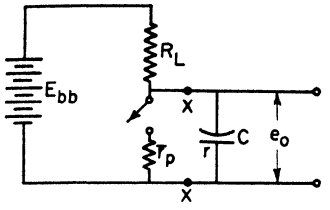


Fig. 1.64. Equivalent plate circuit of the saw-tooth voltage generator of Fig. 1.63.

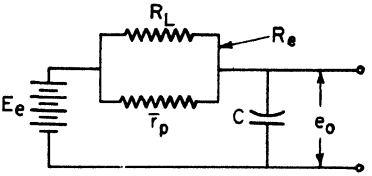


Fig. 1.65. Thevenized equivalent of Fig. 1.64.

The charging time constant is readily obtained as

$$T_{ch} = R_L C$$

but, the discharging time constant must be obtained from the Thevenin's theorem reduction of the circuit shown in Fig. 1.65 where

$$E_e = E_{bb}\left(\frac{\bar{r}_P}{\bar{r}_P + R_L}\right) \tag{1.60}$$

and

$$R_e = \left(\frac{\bar{r}_P R_L}{\bar{r}_P + R_L}\right) \tag{1.61}$$

Consequently, the time constant is

$$T_{dsch} = R_e C \tag{1.62}$$

Hence, during discharge, the output voltage is

$$e_0 = E_e - (E_e - r_1)\epsilon^{-t/R_e C} \tag{1.63}$$

Whereas, during charging

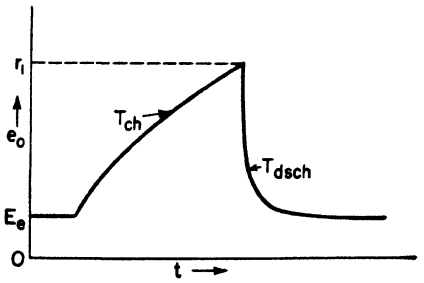


Fig. 1.66(a). Calculated output from saw-tooth generator of Fig. 1.63.

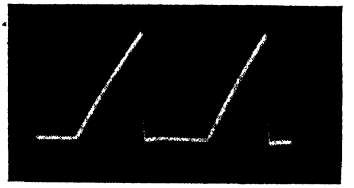


Fig. 1.66(b). Actual output from a saw-tooth generator when the switch tube is gated by a square wave.

$$e_0 = E_{bb} - (E_{bb} - r_2)\epsilon^{-t/R_L C} \tag{1.64}$$

Thus, the waveform appears as shown in Fig. 1.66 and approximates the idealized waveform of Fig. 1.61 as long as

$$T_{ch} \gg T$$
$$T_{dsch} \ll T$$

1.14 Methods of Improving Linearity

Although the saw-tooth voltage waveform obtained in the preceding article is extremely useful in low-cost oscilloscopes, or in other devices where extreme accuracy is not required, many cases arise in high-precision sweep or delay circuits where the slight exponential curvature is highly undesirable and methods of compensation must be employed to produce the necessary degree of linearity.

The most direct method would be to increase the plate supply voltage (E_{bb}) and use a smaller portion of the resulting exponential as illustrated in Fig. 1.67. However, a definite practical limit exists.

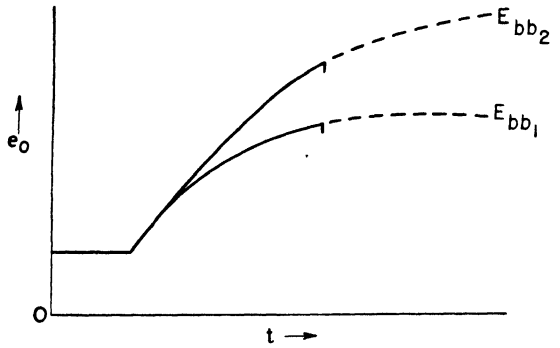


Fig. 1.67. Improvement in linearity by increasing the plate supply voltage.

A more satisfactory, although more complex, method would be to charge the condenser from a constant current source. Since the voltage across the sweep capacitor is given by

$$e_0 = \frac{1}{C}\int i\,dt \tag{1.65}$$

then, if the current is a constant with respect to time,

$$e_0 = \frac{1}{C}\left(It + Q_0\right) \qquad \text{where } Q_0 = \text{initial charge} \tag{1.66}$$

This is the equation of a straight line. With a suitable switch tube, a very linear saw tooth is obtained.

One of the most convenient constant current sources is a pentode vacuum tube connected as a two-terminal circuit element as indicated in Fig. 1.68. Over a restricted range, governed by the tube potentials,

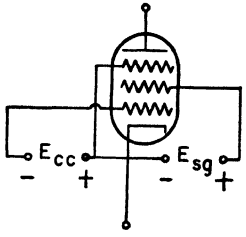


Fig. 1.68. Pentode connections for use as a two-terminal constant-current device.

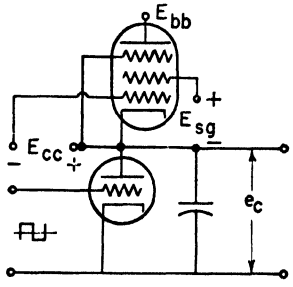


Fig. 1.69. Saw-tooth voltage generator using a pentode as a constant-current source.

the current through the pentode is independent of the voltage across it. Connecting the pentode in place of the switch tube load resistor as shown in Fig. 1.69 provides the required constant-current source.

Another method of obtaining the desired constant-current source is to use feedback as shown in Fig. 1.70. This circuit is commonly called a "bootstrap sweep circuit." Prior to the instant that the switch tube cuts off, the voltage at point (A) is practically the plate supply voltage (E_{bb}) since the drop in the diode will be comparatively small. The other side of the feedback condenser (C_{fb}), which has a very large capacitance, is at a very low voltage. When the switch tube cuts off, the saw-tooth voltage starts developing and is fed to the grid of the cathode follower. The output of the cathode follower

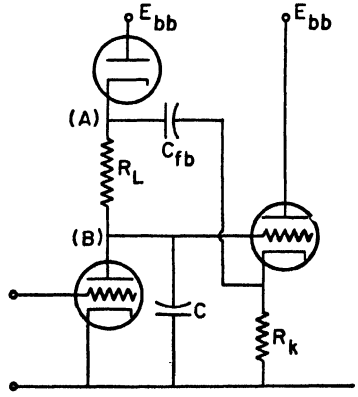


Fig. 1.70. Bootstrap saw-tooth voltage generator.

is then fed back to point (A). Since the feedback condenser is quite large, little or no change in voltage occurs across it during the sweep so that the voltage at point (A) increases correspondingly with the saw tooth. Since the voltage at point (B) *is* the saw-tooth, points (A) and (B) are increasing together so that the voltage across R_L is maintained substantially constant, effectively producing a constant current source. Since the voltage applied by C_{fb} might cause point (A) to rise above E_{bb}, the diode is necessary, as an automatic switch, to insure that this cannot occur. Since the gain of the cathode follower is not unity, but slightly less, and since the voltage across the feedback condenser *will* change slightly, a mild downward curvature of the waveform will result. The circuit may be calculated by conventional methods.

A great many variations in these two basic schemes are available to the designer, and although the details vary somewhat, the principles are basically those enunciated here.

1.15 The Trapezoidal Voltage Generator (Saw-tooth Current)

The saw-tooth voltage generator of Arts. 1.13 and 1.14 is widely used as a sweep-voltage source in electrostatically deflected cathode ray tubes. The inherently superior properties of the magnetically deflected cathode ray tube in most radar and television applications demand the development of a saw-tooth *current* generator since the deflection in such cathode ray tubes is proportional to current rather than voltage. To determine the voltage necessary to produce this current waveform, consider the equivalent circuit of a deflecting coil as shown in Fig. 1.71. Assume that the desired current waveform

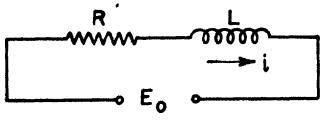


Fig. 1.71. Equivalent circuit of a deflecting coil in a magnetically deflected cathode ray oscilloscope.

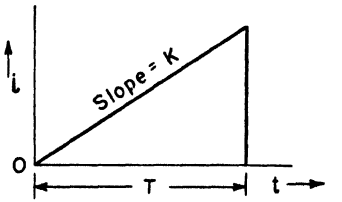


Fig. 1.72. Saw-tooth current.

is as indicated in Fig. 1.72. During the interval of time T, the equation for the current may be written as

$$i = Kt \qquad \text{where } K = \text{slope} \qquad \text{and} \qquad t = \text{time}$$

Now, the Kirchhoff loop equation for the deflecting coil is

$$e = L\frac{di}{dt} + Ri \tag{1.67}$$

but

$$\frac{di}{dt} = \text{slope} = K$$

and

$$i = Kt$$

hence, the expression for the applied voltage must be

$$e = (LK) + (RK)i \tag{1.68}$$

Now, since the slope (K) is a constant, the waveform of e must be as indicated in Fig. 1.73, which is obviously a trapezoid. The

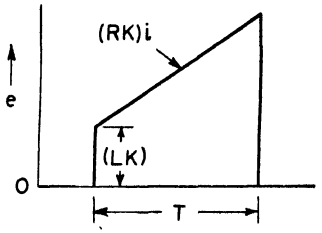


Fig. 1.73(a). Trapezoidal voltage required to produce the sawtooth current of Fig. 1.72.

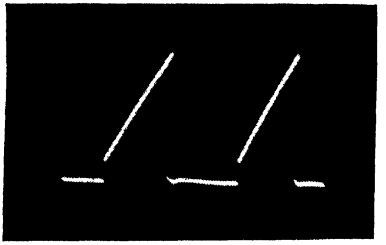


Fig. 1.73(b). Actual output from a trapezoidal voltage generator when the switch tube is gated by a square wave.

quantity (LK) is referred to as the "initial jump" and the slope of the voltage waveform is (RK). The jump-to-slope ratio is

$$\frac{\text{jump}}{\text{slope}} = \frac{L}{R} = \text{time constant of the deflecting coil} \tag{1.69}$$

The development of this voltage in an R-C circuit is not difficult since it consists essentially of a saw tooth superimposed on a d-c value. Such a result may be obtained by splitting the resistance

in the *R-C* saw-tooth voltage generator as shown in Fig. 1.74. The output voltage is given by the expression

$$e_0 = E_{bb} - iR_1 \tag{1.70}$$

but

$$i = \left(\frac{E_{bb} - r}{R_1 + R_2}\right)\epsilon^{-t/(R_1+R_2)C} \tag{1.71}$$

where r is the initial condenser voltage,

so

$$e_0 = E_{bb} - (E_{bb} - r)\left(\frac{R_1}{R_1 + R_2}\right)\epsilon^{-t/(R_1+R_2)C} \tag{1.72}$$

At the initial instant ($t = 0$)

$$e_0 = \text{initial jump} = E_{bb} - (E_{bb} - r)\left(\frac{R_1}{R_1 + R_2}\right) \tag{1.73}$$

and the initial slope is

$$\left.\frac{de_0}{dt}\right|_{t=0} = (E_{bb} - r)\frac{R_1}{(R_1 + R_2)^2 C} \tag{1.74}$$

which yields the desired trapezoidal voltage.

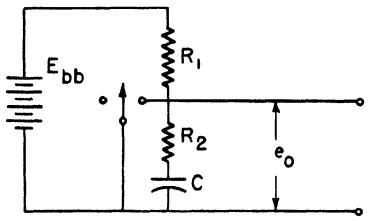


Fig. 1.74. *R-C* trapezoidal voltage generator.

In an actual circuit, the switch would be replaced by the customary switch tube and the details on discharge of the sweep condenser would be somewhat more laborious due to the introduction of the static plate resistance. However, the method of analysis is the same as for the other circuits.

1.16 Review of R-L-C Circuits

The characteristic transient responses of the *R-L-C* circuits find frequent application in wave-shaping circuits and a brief review of their properties is in order before proceeding to the actual circuits. Consider first the series *R-L-C* circuit. The equation for equilibrium, with constant voltage applied, is

$$E_0 = L\frac{di}{dt} + Ri + \frac{1}{C}\int i\,dt \tag{1.75}$$

where E_0 is the impressed voltage.

The solution of this differential equation is standard and may fall into any one of three forms depending upon the relative values of R, L, and C. The three possible solutions are:

(1) Over-damped when $R > 2\sqrt{L/C}$.
(2) Critically damped when $R = 2\sqrt{L/C}$.
(3) Oscillatory when $R < 2\sqrt{L/C}$.

Typical current waveforms for each of these three cases are shown in Fig. 1.75.

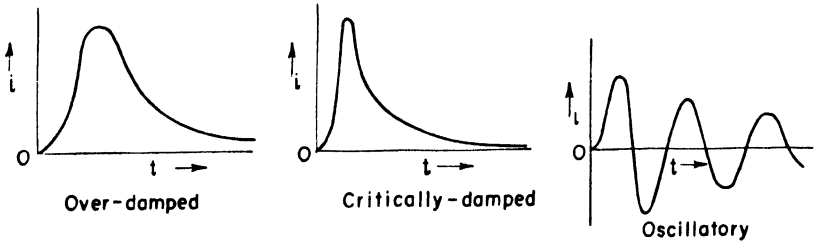


Fig. 1.75. The three possible solutions to the R-L-C circuit equation.

In the lossless case, the resonant frequency is given by

$$\omega_0 = \frac{1}{\sqrt{LC}} \tag{1.76}$$

The Q of the series R-L-C circuit is

$$Q = \frac{\omega_0 L}{R} = \frac{\sqrt{L/C}}{R} \tag{1.77}$$

Hence:

In the over-damped case $Q < \frac{1}{2}$.
In the critically damped case $Q = \frac{1}{2}$.
In the oscillatory case $Q > \frac{1}{2}$.

For a parallel R-L-C circuit supplied by a constant current source, the equation for equilibrium is

$$I_0 = C\frac{de}{dt} + Ge + \frac{1}{L}\int e\,dt \tag{1.78}$$

which solves out as before, yielding voltages of the form obtained

for the currents in the preceding case and under the following conditions:

(1) Over-damped case $R < \frac{1}{2}\sqrt{L/C}$.
(2) Critically damped case $R = \frac{1}{2}\sqrt{L/C}$.
(3) Oscillatory case $R > \frac{1}{2}\sqrt{L/C}$.

The circuit Q is given by

$$Q = \frac{R}{\omega_0 L} = \frac{R}{\sqrt{L/C}} \tag{1.79}$$

Then:

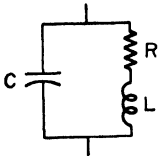


Fig. 1.76. Conventional tuned circuit.

For the overdamped case $Q < \frac{1}{2}$.
For the critically damped case $Q = \frac{1}{2}$.
For the oscillatory case $Q > \frac{1}{2}$.

A similar situation exists for the *R-L-C* circuit of Fig. 1.76 except that the Q is given by

$$Q = \frac{\omega_0 L}{R} \tag{1.80}$$

1.17 R-L-C Peaker

Examination of the waveforms of Fig. 1.75 in the preceding article shows that under the condition of critical damping, or very close to it, the output voltage in a parallel *R-L-C* circuit is peaked. This type of peaking action is sometimes more desirable than that obtained from the conventional *R-C* peaker. The complete circuit of a typical practical peaker is shown in Fig. 1.77. An inductance is

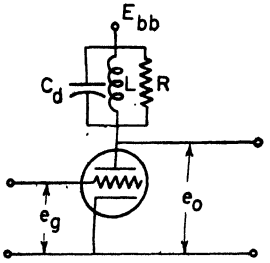


Fig. 1.77. *R-L-C* peaker circuit.

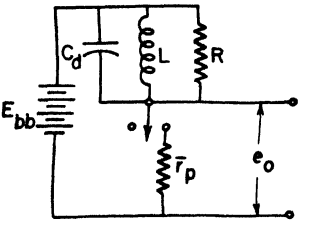


Fig. 1.78. Equivalent plate circuit of the *R-L-C* peaker.

usually placed in the plate circuit of the switch tube and is then paralleled by a resistance which is adjusted so that

$$R = \tfrac{1}{2}\sqrt{L/C_d} \tag{1.81}$$

where C_d is the distributed capacitance.

The usual equivalent circuit is shown in Fig. 1.78. While the tube is conducting, a steady current flows through the inductance. When the tube is cut off, this current is interrupted and the typical critically damped voltage waveform appears in the output. A similar pulse is generated when the switch tube is cut back on again except that the additional series resistance due to the tube lowers the overall Q, reducing the amount of peaking as illustrated in Fig. 1.79.

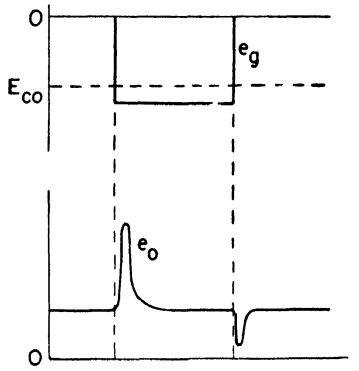


Fig. 1.79. *R-L-C* peaker waveforms.

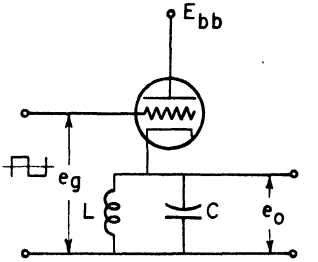


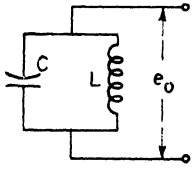
Fig. 1.80. Ringing circuit.

1.18 The Ringing Circuit

The ringing circuit is essentially the same as the *R-L-C* peaker, except that the *R-L-C* circuit is tuned to resonance, rather than being critically damped. Although the tuned circuit may be located in either the cathode or the plate circuit, it is more generally placed in the cathode circuit due to the very desirable damping introduced by cathode follower action during tube conduction. The reason for this may be more apparent subsequently.

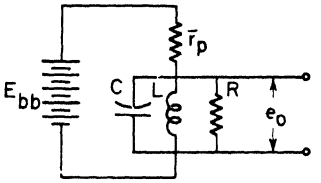
For the purposes of discussion, assume a switch tube with a tuned *L-C* circuit in its cathode as shown in Fig. 1.80. Assuming the tube to be conducting, then in the steady state, a constant current

flows through the inductance. When the tube cuts off, the circuit oscillates, or "rings." The appropriate equivalent circuits are shown in Fig. 1.81. Thus, when the tube is nonconducting, the circuit has



Equivalent Circuit of Ringing Circuit when Tube is cut-off

$Q = \frac{\omega L}{R}$



Equivalent Circuit of Ringing Circuit during Tube conduction

$R = \frac{r_p}{\mu + 1}$ and $Q = \frac{R}{\omega L}$

Fig. 1.81. Ringing circuit equivalent circuits.

a high Q and rings with very little damping. However, during tube conduction, the low output resistance of the tube which is then working as a cathode follower shunts the tuned circuit, producing a very low Q, damping out the oscillations with great rapidity. As a result, the waveform is as indicated in Fig. 1.82.

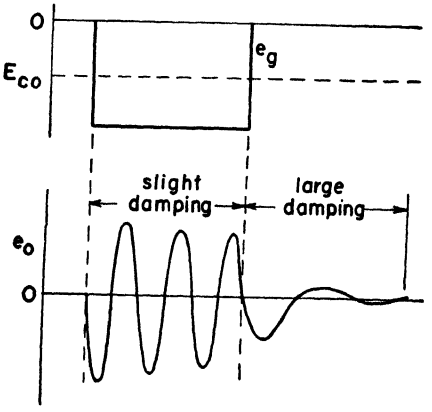


Fig. 1.82. Ringing circuit waveforms.

1.19 Effects of Distributed Capacitance

In the preceding treatment the effects of wiring capacitances and tube interelectrode capacitances have been neglected. Inasmuch as

most shaping circuits develop, or involve, waveforms with sharp edges and discontinuities, it is apparent that the high-frequency harmonic content must be large. The effect of these shunt capacitances is to attenuate the high-frequency terms, resulting in:

(1) Broadening of narrow pulses.
(2) Reduction of narrow pulse amplitude.
(3) Delay of steep wavefronts slightly.

The theoretical analysis may be made more exact, albeit more cumbersome, by including these shunt capacitances in the appropriate equivalent circuits. Although they may frequently be neglected, if time intervals as large as 10 to 50 μ secs are important, then the distributed capacitances must be considered if the theoretical analysis is to coincide with practical results.

As a consequence, construction of wave-shaping circuits should be carefully done to avoid excessive wiring capacitances, while the proper selection of tubes and the use of pentodes wherever it is economically feasible will provide a further reduction. Nevertheless, some stray capacitances will always be present in an undetermined amount, and exact correlation of experimental and calculated waveforms cannot be expected.

In most cases, but not always, from the engineering viewpoint, the increase in accuracy obtained by mathematical consideration of the distributed capacitances does not warrant the expenditure of time and effort necessary to reach a solution. Generally, the circuits are designed and built by the methods outlined thus far, and the calculated circuit values are then juggled experimentally to obtain, as nearly as possible, the precise results desired. A few special cases exist where consideration of the stray capacitances does not introduce severe mathematical difficulties (e.g., the pulse amplifier), but such circuits are definitely the exception rather than the rule.

1.20 Artificial Lines as Pulse Generators and Delay Circuits

The characteristics of transmission lines and their responses to sinusoidal signals are generally covered in the usual introductory courses, and are reviewed in Chap. 4 of the text. This article is concerned, in a qualitative way, with the response of artificial lines to pulses and static charging.

The artificial line is a common laboratory tool used to provide a convenient equivalent representation of an actual line in a limited

amount of space. In general, such a line consists of a number of T, π, or L filter sections connected in tandem, the number of sections used being principally contingent upon the extent to which simulation of the real line is required. The greater the number of sections per wavelength, the more accurate the representation.

Associated with such lines is a definite velocity of propagation, characteristic impedance, attenuation, and phase shift constant for any given frequency. By reducing the losses in the reactance elements to the point where they are negligible, and by making $L/C \gg \omega^2 L^2/4$, the attenuation becomes negligible and, within a fair degree of approximation, the dependence of velocity and characteristic impedance upon frequency is removed. The phase shift becomes a linear function of frequency. On such a line, all frequencies travel with about the same velocity, have their phases retarded proportionally, and encounter roughly the same characteristic impedance. Thus, Z_c is the characteristic impedance, $\sqrt{L/C}$, and V the velocity, $1/\sqrt{LC}$, in sections per second, where L is the inductance per section in henries and C the capacitance per section in farads.

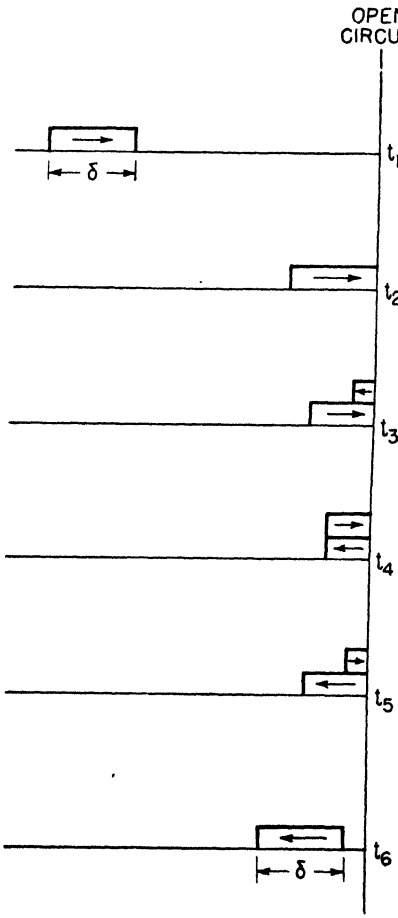


Fig. 1.83. Life history of a rectangular voltage pulse on an open-circuited lossless line.

Let N be the number of sections. Then, the time required for a wave to travel the length of the line is T, where

$$T = \left(\frac{1}{V}\right)N \qquad (1.82)$$

so

$$T = N\sqrt{LC} \qquad (1.83)$$

Consider the case of a pulse applied to an open-circuit lossless artificial line on which the duration of the pulse is such that it does not contain any significant frequency components outside the pass band of the tandem filters. The life history of such a pulse could be depicted as shown in Fig. 1.83. Since the pulse travels the length of the line, is reflected, and returns to the sending end, its total time of travel is $2T$ or

$$\tau = 2N\sqrt{LC} \tag{1.84}$$

which is the length of time that the pulse has been delayed with respect to the time that it was first placed on the line. Hence, in such an application, it is usually called a "delay line."

A somewhat different application results from static charging of the line, producing pulses of certain duration. The line is then called a "pulse-forming line." Consider a statically charged line as shown in Fig. 1.84. Such a line actually has two voltage components on it, as shown, in order to meet the terminal conditions of zero current imposed by the open circuit. The condition corresponds to time t_4 in Fig. 1.83 for that section of line which is charged. After the line has been charged, as indicated, if the open end of the line were closed through the characteristic impedance of the line, the line would discharge itself in the length of time required for both waves to traverse the line. Hence, the pulse length, generated across this termination, would be twice the time required for the pulse to traverse one length of the line, or

$$\delta = 2N\sqrt{LC} \tag{1.85}$$

In the ideal case the waveform would appear as shown in Fig. 1.85(a).

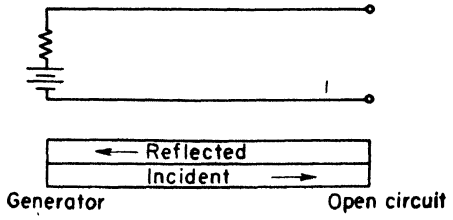


Fig. 1.84. Static charging of an open-circuited lossless transmission line.

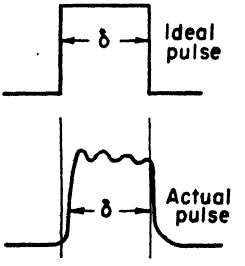


Fig. 1.85. Ideal and actual waveforms for a pulse-forming line.

Actually, due to reflections between sections, it is more like Fig. 1.85(b), the irregularities becoming smaller and more numerous as N is increased. In certain applications, the generated pulse is not used directly and the irregularities may be removed by applying it to an ordinary overdriven amplifier. In such cases, 2 to 5 sections are usually adequate. In applications where the pulse must be used directly, large degrees of irregularity cannot be tolerated and anywhere from 6 to 16 sections are commonly used.

PROBLEMS

1.1 A trapezoidal voltage wave of the form shown in Fig. 1.86 is applied to the horizontal deflecting plates of a cathode ray tube through a conventional coupling circuit. Calculate the amount of d-c bias that must be added to the coupling circuit to center the waveform on the CRT if

(a) $o - a = 1000\ \mu$ sec $\quad a - b = 0\ \mu$ sec $\quad b - c = 1000\ \mu$ sec.

(b) $o - a = 500\ \mu$ sec $\quad a - b = 500\ \mu$ sec $\quad b - c = 1000\ \mu$ sec.

(c) $o - a = 200\ \mu$ sec $\quad a - b = 800\ \mu$ sec $\quad b - c = 1000\ \mu$ sec.

(d) Design a clamper circuit that will do the same job as the variable bias source.

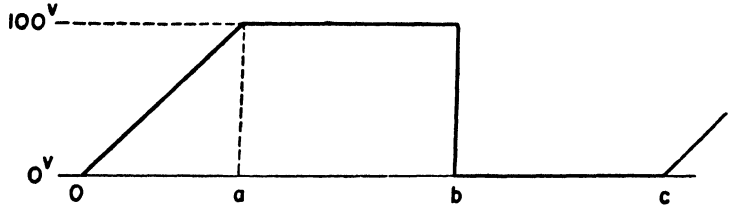


Fig. 1.86. Trapezoidal voltage waveform.

1.2 The two halves of a 6SN7 tube are connected as a two-stage amplifier, but the grid biases are to be adjusted so that they operate as plate current cutoff clippers. For both triode sections, $E_{bb} = 200$ volts and $R_L = 20$K. Assuming perfect interstage coupling, calculate the biases required to convert a 1.0-volt (rms), 1000-cps sine wave into a flat-topped, symmetrical trapezoid which reaches its maximum value in 100 μ sec.

1.3 The output voltage from the second triode clipper of Problem 1.2 is fed into an R-C peaker circuit. Assume that the rise and drop in voltage from minimum to maximum, and conversely, are linear with respect to time, a good practical approximation.

(a) Sketch the output voltage waveform from the peaker.

(b) What time constant should be used to give a maximum signal amplitude of 1 volt from the peaker?

1.4 A 100-cps, symmetrical square wave has an overall peak-to-peak height of 30 volts. It is to be narrowed down to pulses 1 millisec long and spaced at intervals of 10 millisec. The circuit consists of two biased triode stages connected together by an ideal coupling circuit. The plate load resistors for both tubes are 20 K and the power supply voltage is 200 volts. The input coupling circuit to the first tube consists of a 0.01-μf coupling condenser and a 200-K grid leak. This circuit functions as the peaker. 6J5 tubes are used.

(a) Calculate the biases which should be used on each tube if the output pulses are to be positive and 50 volts high.

(b) How long would the crest of the pulse be if the base is the required 1 millisec.

1.5 Two triode amplifiers are coupled together by an *R-C* peaker circuit. The input coupling to the first triode is assumed to be ideal and all of the tube and wiring capacitances can be neglected. The input wave consists of 50-volt negative pulses, 100 μ sec long and 400 μ sec apart. The circuit is to produce an output waveform consisting of narrow positive pulses. The following data applies:

$R_{L_1} = 45$ K	$R_{L_2} = 45$K
$E_{bb_1} = 100$ volts	$E_{bb_2} = ?$
$\mu_1 = 5$	$\mu_2 = 5$
$\bar{r}_g = 2000$ ohms	$\bar{r}_g = 2000$ ohms

$i_{b_1} = i_{b_2} = 2\ ma$ when grid voltage and condenser current are zero

The peaker constants are $C = 100\ \mu\mu$f and $R = 100$ K.

(a) Sketch the waveforms of e_{b_1}, i_{b_1}, and e_{c_2}. Indicate the time constants of all exponential curves and the maximum and minimum values of each waveform.

(b) What is the lowest value of E_{bb_2} permissible if the waveform of e_{b_2} is to be sharp-pointed pulses instead of flat-topped ones?

1.6 A sinusoidal voltage 250 volts in amplitude is applied to a cathode follower using a 6J5 tube with

$R_k = 30$ K	$\bar{r}_g = 1000$ ohms for $e_c \geqq 0$
$E_{bb} = 300$ volts	$R_c = 0.75$ megohm

where R_c denotes the grid-clipping resistor.

(a) Calculate grid and cathode voltage waveforms, referred to ground.

(b) Calculate the circuit gain from the equivalent circuit.

(c) Why is the ratio of the total change in e_k to the total change in input voltage, as determined from the waveforms of part (a), different from the gain calculated in part (b)?

1.7 Design a saw-tooth voltage generator using a 6SN7 tube and a plate supply of 300 volts, which will produce a saw tooth 1000 μ sec long having

a minimum value of 12.5 volts and a maximum value of 65 volts. A 50-volt, 500-cps square wave is available for control of the switch tube. Find R and C.

1.8 In the R-C trapezoidal voltage generator of Fig. 1.74, if

$R_1 = 150$ K $\quad C = 0.005$ μf $\quad \bar{r}_p = 10$ K

$R_2 = 10$ K $\quad E_{bb} = 120$ volts

and if the switch tube is cut off by a 200-μ sec negative gate, draw the waveform of the output voltage and calculate all critical points.

1.9 An R-C trapezoidal voltage generator supplies the grid excitation for a sweep amplifier tube which has a sweep coil in its plate circuit. The switch tube in the trapezoidal voltage generator conducts for 200 μ sec and is cut off for 300 μ sec. The current through the sweep coil is to change by 90 ma during the sweep. If the sweep amplifier tube has a $\mu = 10$ and $r_p = 4000$ ohms, and if the trapezoidal voltage generator of Fig. 1.74 has

$R_1 = 500$ K $\quad C = 0.001$ μf $\quad \bar{r}_p = 10$ K

$R_2 = 10$ K $\quad E_{bb} = 100$ volts

(a) determine the required dimensions of the trapezoidal voltage applied to the grid of the sweep amplifier.

(b) What must be the L and R of the sweep coil for proper operation?

1.10 One half of a 6SN7 is used in a ringing circuit having $E_{bb} = 120$ volts. Assume that the tube is controlled by an external source which cuts the tube off when "ringing" is desired.

(a) If $L = 50$ mhy, what value of C is required for the circuit to ring at a frequency of 9310 cps?

(b) What is the magnitude of the first peak of cathode voltage? Is it positive or negative? Why? (Assume the output current is zero and neglect losses in the tuned circuit.)

(c) What is the smallest magnitude of negative voltage applied from grid to ground of the switch tube which will cut the tube off completely and allow the circuit to ring properly?

(d) When the switch tube cuts on again, calculate the Q of the cathode circuit.

1.11 A constant frequency, variable negative gate width square wave is used in a particular timer. It is clamped at 0 volts by a diode clamper. Draw the circuit diagram of the clamper and the clamped waveform.

CHAPTER 2

TRIGGER CIRCUITS

THE PRECEDING CHAPTER has dealt with circuits that are operated like a doorbell push button, being held open or closed for intervals of time dictated by some external agency, and then combining this action with the forming or shaping properties of simple *R-C*, *R-L*, *R-L-C* circuits. A distinctly different grouping of circuits, widely used in wave-forming and allied operations, will be discussed in this chapter. The operation of these circuits is closely analogous to that of a mousetrap, in that a small triggering impulse causes it to switch from one stable condition to another in a length of time and in a manner largely independent of the nature of the triggering impulse. The circuits then reset themselves and await another similar impulse. Although the tubes operate essentially as switches, the control of the switching interval is internal, rather than external, and it is to be expected that the principle of operation will be unlike that of the circuits of Chapter 1. The "triggering" depends upon the development of a negative resistance characteristic.

Trigger circuits are extensively used in various forms in radar, television, and electronic computing circuits of all types where counting, sorting, multiplication, division, and other such applications are required. Together with the wave-shaping circuits of Chap. 1, they form a vast field of application which, even in the light of present knowledge, is limited principally by the imagination of the engineer. It is one of the most rapidly developing fields of electronic circuit theory.

2.1 The Theory of Trigger Circuits

Trigger circuits are defined as circuits which have two stable conditions of operation for fixed values of circuit constants and applied voltage. The triggering action arises from the fact that the currents and voltages in such circuits can be made to change from one condition of stable equilibrium to the other by adjusting some circuit

constant or operating potential, and can then be made to switch back to the original stable condition by adjusting the circuit element or voltage to some *other value.*

In order to produce the triggering action it is necessary to fabricate a device which has two points of stable equilibrium separated by a region of unstable equilibrium. Considering voltages and currents, such conditions of operation could be obtained by means of a device which generates a current-voltage characteristic with two regions where it is a normal positive resistance, separated by a region of negative resistance.

Negative resistance characteristics may be generated in a number of ways, but they are usually classified into two broad general categories, depending upon whether the current or voltage is the source of control. Hence, the two classifications are

(1) Current-controlled.
(2) Voltage-controlled.

Consider the voltage-controlled case first. A schematic diagram is shown in Fig. 2.1 which resembles a voltage-controlled feedback

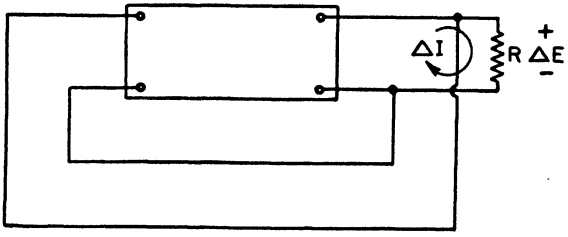


Fig. 2.1. Schematic diagram of a voltage-controlled negative-resistance device.

amplifier. To produce a negative resistance at the terminals of the device, the slight changes in voltage fed back through the device must arrive back at the output in phase with the original disturbance in order to obtain reinforcement. Assume a small increment of voltage of the polarity shown in Fig. 2.1. If connections are such that this voltage, after passing through the device, produces an increment of current ΔI in the direction shown, then this current originates a voltage drop which adds to the original incremental voltage ΔE. If this voltage drop, due to the increment in current,

call it ΔE_x, is greater than the original incremental voltage assumed, ΔE, which produced ΔI, then a negative resistance has been produced since the original increment has been regenerated. For instability to exist, the total resistance across the output terminals must be negative. The device itself may be represented by a negative resistance, and the equivalent circuit appears as shown in Fig. 2.2. Now, the total resistance R' is

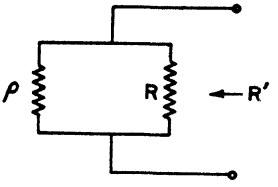


Fig. 2.2. Total output impedance of the circuit of Fig. 2.1.

$$R' = \frac{R\rho}{R + \rho} \tag{2.1}$$

but
$$\rho = -|\rho| \tag{2.2}$$

so
$$R' = \frac{-R|\rho|}{R - |\rho|} \tag{2.3}$$

or
$$R' = \frac{R \cdot |\rho|}{|\rho| - R} \tag{2.4}$$

Hence, in order for the total resistance across the output terminals (R') to be negative, the load resistance (R) must be greater than the absolute value of the negative resistance.

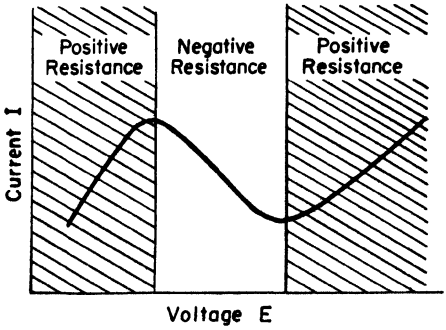


Fig. 2.3. Typical voltage-controlled current—voltage characteristic.

This same result could be deduced directly from the characteristic current-voltage curves which, for the voltage-controlled case, have the general form shown in Fig. 2.3. Assume a value of load resistance

(R) which is less than the magnitude of the negative resistance. Draw the load line on the current-voltage characteristic in the same manner as for a vacuum tube. The result appears as shown in Fig. 2.4.

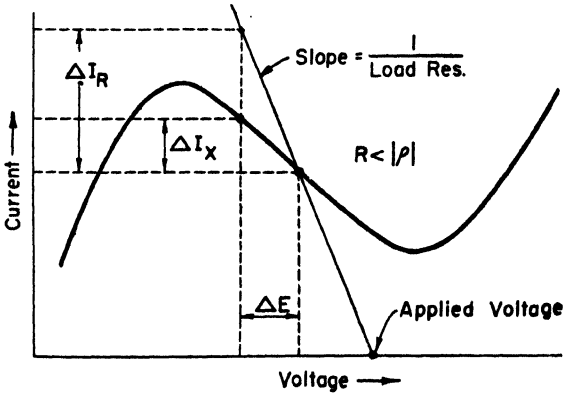


Fig. 2.4. Conditions for stable operation of a voltage-controlled negative resistance device.

Assume a change in voltage across the load of ΔE, as shown. A current change of ΔI_R would have been required to produce ΔE (that is, to regenerate it), but this change in voltage applied to the device only produces an incremental change in current of ΔI_z, which

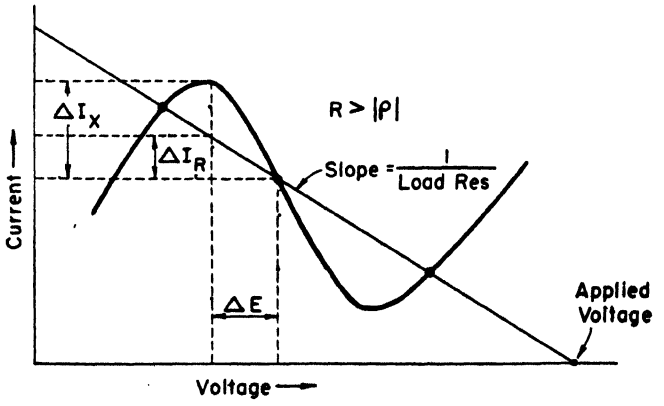


Fig. 2.5. Conditions for unstable operation of a voltage-controlled

is obviously less than the ΔI_R required for regeneration. Therefore, this is a *stable* condition of operation.

However, suppose we assume that the load resistance has a value which is larger than the absolute value of the negative resistance, then the load line appears on the current-voltage characteristic as shown in Fig. 2.5. Following the same method as before, assume an incremental reduction in voltage of ΔE. From Fig. 2.5 we see that a current change of ΔI_R is the minimum required to generate ΔE by passing through R, the load. Now, since the device actually produces an increment of current ΔI_z, which is greater than the ΔI_R required, it follows that regeneration of the original increment occurs, resulting in instability. Thus, we arrive at the same conclusion that, in a voltage-controlled negative resistance device, instability occurs when the load resistance is greater than the absolute value of the negative resistance generated by the device. A similar analysis applied at the other two possible operating points indicated in Fig. 2.5 will indicate that they are points of stable operation.

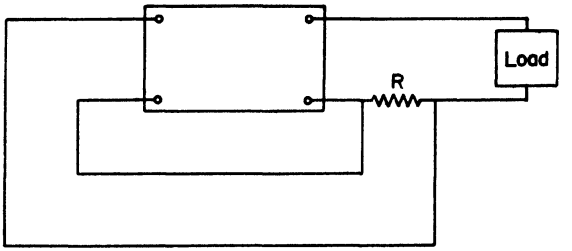


Fig. 2.6(a). Schematic diagram of a device for generating current-controlled negative resistance.

Current-controlled negative resistance is developed as shown in Fig. 2.6(a) where the feedback voltage is proportional to the load current. A typical current-controlled current-voltage characteristic is shown in Fig. 2.6(b). It may be shown by exactly the same method as that employed for the voltage-controlled case that instability results when the

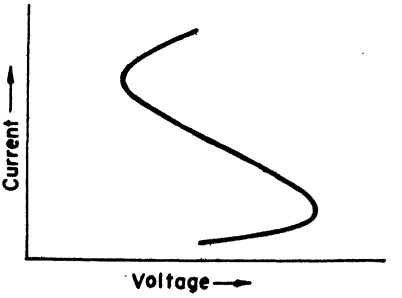


Fig. 2.6(b). Typical current-controlled current-voltage characteristic.

load is less than the absolute value of the negative resistance generated.

Considering the voltage-controlled characteristic again, it is apparent that if the load line were progressively shifted, parallel to itself, by increasing the applied voltage E_b, a point would be reached where the load line is just tangent to the current-voltage characteristic, and operation would be in the unstable region. Any slight change in voltage across the load resistance R would cause the point

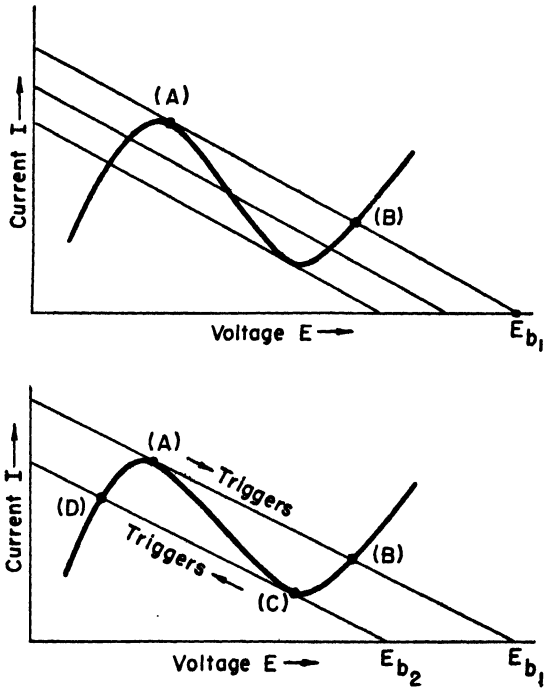


Fig. 2.7. Triggering a voltage-controlled device by changing the supply voltage.

of operation to jump from point (A) to point (B) in a very short period of time, and we say that the circuit has been "triggered." Conversely, if we reduced E_b to a point E_{b_2} where the load line was tangent to the bottom bend in the characteristic, triggering would again occur from point (C) to point (D). This is illustrated in Fig. 2.7.

Triggering may also be accomplished by changing the slope of the load line instead of its position, by varying the size of the load resistance. This is illustrated in Fig. 2.8. Still another possible method of triggering would be to change the location or shape of the characteristic curve by changing one of the electrode potentials in a vacuum tube. Consequently, in summary, triggering may be accomplished by:

(1) Variations in the supply voltage.
(2) Variations in the load resistance.
(3) Variations in electrode potentials.

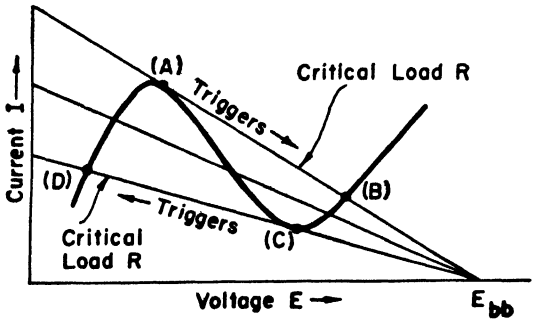


Fig. 2.8. Triggering a voltage-controlled device by changing the load resistance.

The classification according to the method of control of the negative resistance is one method of separation of these deviees, but another is based upon the manner of generation of the negative resistance characteristic, as follows:*

(1) Those devices in which the element exhibiting the negative resistance also controls the internal phenomenom which makes the negative slope possible. The negative resistance is part of the d-c characteristics (e.g., the dynatron oscillator).
(2) Those devices in which the element controlling the internal phenomenon must be directly connected to the circuit of the controlled element before negative resistance is possible. The negative resistance is part of the d-c characteristic. (e.g., the negative transconductance oscillator).

*See Herold, I. R. E., Oct. 1935.

(3) The reverse-phase coupled devices which depend upon phase shift in the feedback circuit and in which the controlling element is separate from the controlled. The negative resistance is part of the a-c characteristics only. (e.g., most conventional oscillators such as the Hartley, Colpitts, and others).

2.2 Development of Negative Resistance in Tetrodes

Due to secondary emission, the plate characteristics of screen grid tetrodes are similar, over a confined region, to the characteristic curves of Fig. 2.3 for the voltage-controlled negative resistance device. Hence, by inserting a suitable load resistance in the plate circuit, it could be made to function as a trigger circuit. Unfortunately, the slope of this portion of the tetrode characteristic is usually so low, that is, the negative resistance so very large, that very high values of load resistance must be used to produce instability, thus necessitating high supply voltages. Furthermore, the shape of the characteristic varies somewhat with age, making constant predictable operation difficult. Consequently, the tetrode has found little application as a trigger circuit.

2.3 The Pentode Trigger Circuit

When the suppressor is connected to the screen grid of a conventional pentode in such a way that a change in screen voltage is accompanied by a proportional change in suppressor voltage, the screen current-screen voltage characteristic appears as shown in Fig. 2.9.

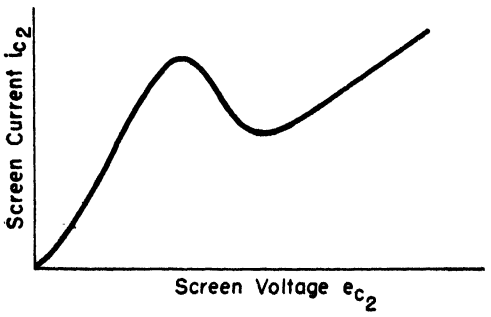


Fig. 2.9. Typical tetrode characteristic showing the negative resistance effect.

The existence of this conformation in the characteristic is not difficult to justify. The electrode potentials are adjusted so that a virtual cathode is formed between the screen and suppressor grids. Under these conditions, a fraction of the current drawn from the cathode is returned to be collected by the screen. Since the screen and suppressor are connected, a decrease in screen voltage, which would ordinarily tend to reduce the screen current, also reduces the suppressor voltage proportionately, making it more negative and causing more electrons to be returned to the screen, thus *increasing* the screen current. Conversely, an increase in screen voltage increases the suppressor voltage, allowing more electrons to pass through to the plate, reducing the number going to the screen, thus reducing the screen current. The magnitude of the negative resistance effect so generated is readily controlled by means of the control grid voltage.

Since the shape of the characteristic is essentially the same as that of the basic voltage-controlled negative-resistance device, the insertion of a load resistance in the screen circuit, and providing

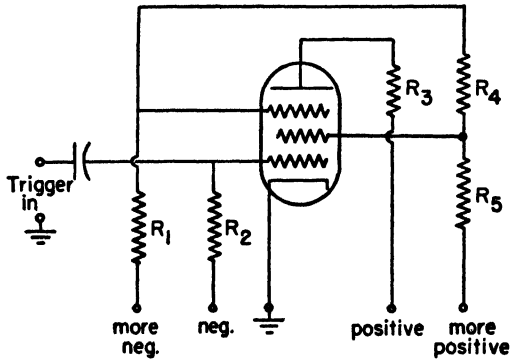


Fig. 2.10. Pentode trigger circuit.

proper coupling between screen and suppressor, will yield a trigger circuit. The circuit diagram appears as shown in Fig. 2.10. The circuit can be triggered by any of the methods previously listed.

2.4 The Eccles-Jordan Trigger Circuit

The Eccles-Jordan trigger circuit, originally devised in 1919, is probably the most useful one yet contrived, and the one upon which

most subsequent development has been based. The circuit diagram of the basic Eccles-Jordan circuit is shown in Fig. 2.11. The triggering action arises since there are two stable conditions of operation:

(1) Tube (1) conducting and tube (2) cut off,
(2) Tube (1) cut off and tube (2) conducting,

separated by an unstable condition which exists when both tubes are conducting. To prove that such a condition is unstable, assume that both tubes are conducting equal amounts simultaneously. If, due to

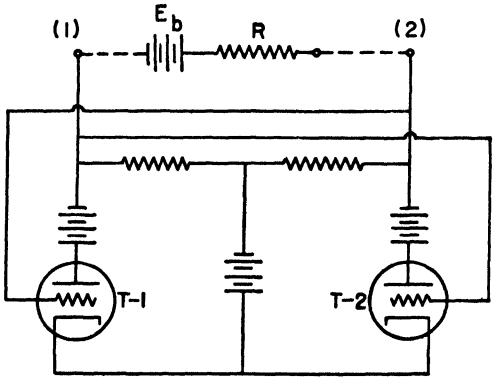


Fig. 2.11. Basic Eccles-Jordan trigger circuit.

any transient circumstance, the current through either tube increases, the drop across the associated coupling resistor increases. This causes the grid of the other tube to go more negative, reducing its plate current and causing the drop in its coupling resistor to be decreased. This raises the voltage on the grid of the tube in which the original change occurred, further increasing the plate current. The action is cumulative and ends when the other tube is cut off.

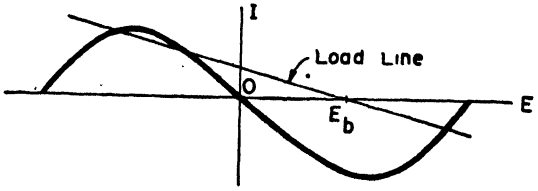


Fig. 2.12. Eccles-Jordan current-voltage characteristic.

The application of a battery and a resistance in series across terminals (1) and (2) results in a current-voltage characteristic of the form shown in Fig. 2.12. Consequently, the triggering action could also be predicted from an analysis of this curve. It is apparent that triggering can be produced by any of the usual methods previously outlined.

As far as the circuit operation is concerned, the tubes merely function as switches, as discussed in Chap. 1, switching between two separate conditions of stable equilibrium under the dictate of some triggering signal.

2.5 Hard-Tube Trigger Circuit

The practical form of the Eccles-Jordan trigger circuit is shown in Fig. 2.13 where "hard," or vacuum, tubes are used. The term is

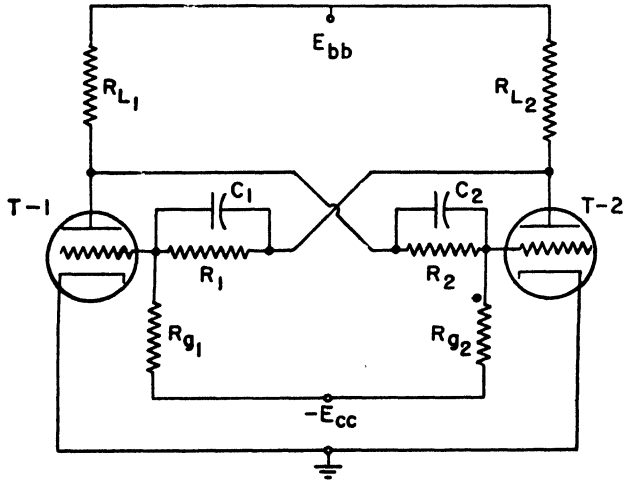


Fig. 2.13. Practical Eccles-Jordan trigger circuit.

applied in order to differentiate between vacuum (hard) tubes and gas (soft) tubes. The circuit is essentially the same as the basic Eccles-Jordan except that only one power supply is required. The switching action is exactly the same.

Assuming the system to be in stable equilibrium when T-1 is conducting and T-2 is off, the circuit can be triggered over to the reverse situation by introducing additional negative voltages in the grid or

plate circuit of T-1, or positive voltages on T-2, by decreasing R_{L_1}, or by varying R_1 or R_{g_1}. Of course, the converse is true if T-1 is off and T-2 is conducting. A common method of triggering is to apply voltage pulses to both grids, or plates, at the same time. Under such conditions, the circuit is sensitive to either positive or negative impulses.

The mechanism by which triggering occurs when voltage impulses are applied to the plates or grids is worth additional consideration. The two cases are essentially the same due to the coupling between grids and plates, so that discussion of either method will suffice for the other. For the moment, neglect the tube interelectrode capacitances as well as the coupling condensers C_1 and C_2. Then, if a negative impulse is applied to both grids, assuming T-1 is on and T-2 off, nothing further happens directly to T-2 since it is already cut off. However, a positive pulse appears at the plate of the conducting tube, T-1, and is coupled over to the grid of T-2 tending to nullify the effect of the original negative pulse on the grid. As long as the amplification of T-1 exceeds the voltage divider ratio in the coupling circuit, the positive pulse produced will obscure the negative pulse applied, yielding a net positive impulse. If the original trigger was large enough for this net pulse to bring the grid of T-2 above cutoff, the usual cumulative switching action will occur.

The reliability of the triggering action is reduced due to the inevitable presence of the tube interelectrode capacitances. Normally, these capacitances tend to stabilize the circuit since they act to prevent sudden changes in electrode voltages. In other words they tend to short out the triggering impulse, that is, to by-pass it. This means, therefore, that instead of having the output pulse from the conducting tube simply divided by the coupling circuit and appearing in reduced amplitude on the grid, the pulse is effectively shorted to ground, and the triggering impulse hardly appears at all on the grid. This makes triggering unreliable and insensitive. As a consequence, it becomes necessary to insert the coupling condensers C_1 and C_2, which then produce a capacitance voltage divider in addition to the resistive one already present. Since the voltage across a capacitive circuit divides in inverse proportion to the capacitances, if we make the coupling condensers relatively large compared to the tube interelectrode capacitances, then they are effectively short circuits, compared to the tube capacitances, during the appli-

cation of the trigger pulse. Hence, the bulk of the trigger appears directly on the grid. This provides extremely close coupling between tubes, considerably enhancing the reliability of triggering. For proper operation, the duration of the triggering impulse must be short compared to the discharge time constant of the coupling condenser. Hence, for high-frequency operation, the discharge time constant, which is necessarily small, requires that extremely short triggering impulses be used.

Since the action of the coupling condensers compensates for the effect of the interelectrode capacitances, in the analysis of the equivalent circuit, to a reasonable approximation, all capacitances can be neglected. In any case, the net effect of these capacitances is to cause a slight rounding of the edges of the steep wavefronts generated. Following the method of Chap. 1, the circuit can be reduced to its equivalent circuit by replacing the tubes with resistors and switches, as shown in Fig. 2.14. The switches operate in synchronism.

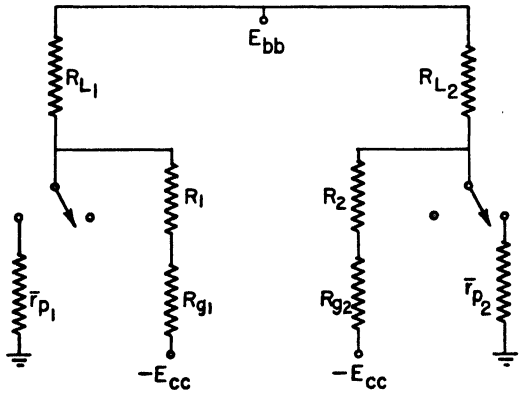


Fig. 2.14. Approximate equivalent circuit of Eccles-Jordan trigger circuit.

2.6 Gas-Tube Trigger Circuit

The gas- or soft-tube trigger circuit, while depending in general upon the same principle of operation as the hard-tube trigger circuit, goes through its triggering cycle in a different manner. A gas-filled triode, such as a thyratron, is equivalent to an electrical switch, in some respects, like the vacuum tube. However, the difference exists due to the fact that the grid has practically no control over con-

duction once the tube starts passing current, and the voltage drop across the tube is practically independent of the tube current.

Assuming the grid to be sufficiently negative when plate voltage is applied, no plate current flows. As the grid voltage is reduced, a certain critical point is reached where the tube conducts and breaks down into a self-sustaining arc discharge over which the grid exerts no control. The current will continue to flow until the drop across the tube is made to fall below a certain critical value long enough for the gas to de-ionize, thus extinguishing the arc and allowing the grid to regain control.

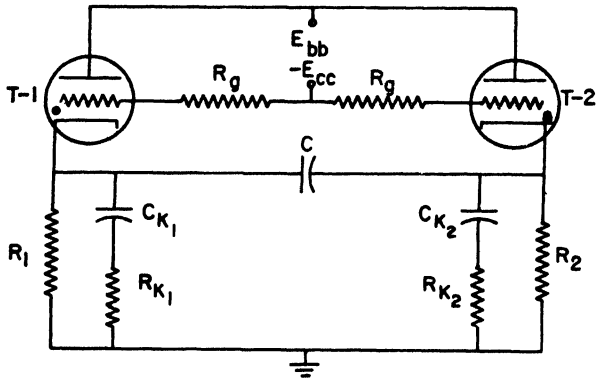


Fig. 2.15. Gas-tube trigger circuit.

To explain the triggering action, assume that in Fig. 2.15 that tube T-1 is conducting, and T-2 is nonconducting. Furthermore, assume that the grid of T-1 is positive and the grid of T-2 is negative. Under this condition, both capacitors C and C_{k_1} are charged to a voltage equal to the drop across R_1. On the other hand, C_{k_2} is uncharged since T-2 is cut off and no current flows through R_2.

Now, if the grid voltage of T-2 is made less and less negative, the critical ignition voltage is eventually reached and the tube arcs over into heavy conduction, abruptly raising its cathode potential up from ground to a value equal to the drop produced across R_2 by the plate current in T-2. The side of capacitor C that is connected to T-1 is already at a voltage equal to the drop across R_1, and when T-2 conducts, both sides of the condenser jump up in potential by an amount equal to the drop across R_2. Consequently, at this instant,

the cathode voltage on T-1 is $(V_{R_1} + V_{R_2})$. For proper operation, this voltage must be greater than the plate voltage on T-1 and must remain greater long enough for the tube to de-ionize completely. Hence, the time constant of the discharging circuit must be large enough to meet this condition. Furthermore, in order for the grid of T-1 to regain control, the grid voltage must have been made more negative as the grid of T-2 is made more positive so that it will not permit conduction after de-ionization.

Condenser C completely discharges and then recharges again in the opposite direction in a short interval of time. When T-1 cuts off, its plate current becomes zero and the voltage across R_1 drops to zero, but the latter is delayed by the discharging and charging currents associated with C and the discharge current of C_{k_1}. The circuit is then in its second stable condition and awaits the application of another triggering impulse. The waveforms appear as shown in Fig. 2.16.

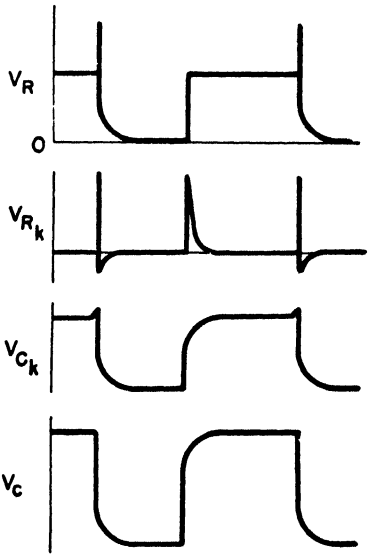


Fig. 2.16. Gas-tube trigger circuit-waveforms.

2.7 Free-Wheeling Plate-Coupled Multivibrator

The trigger circuits considered so far have had the characteristics of a mousetrap, i.e., they can be tripped off in one direction by some externally applied source, then they must be reset by another trigger impulse. However, they do not *self-trigger*. Some external agency is required to fulfill this function.

If, in the hard tube trigger circuit of Fig. 2.13, the resistors R_1 and R_2 are omitted, and the magnitude of the coupling condensers C_1 and C_2 increased, the resulting circuit appears as shown in Fig. 2.17. The difference between this circuit and the Eccles-Jordan is that d-c coupling is no longer used (R_1 and R_2 omitted), having been replaced by a-c coupling. Under such conditions the circuit is capable of self-

triggering so that it oscillates, or "free wheels," at some frequency determined by the network itself, rather than some external source. Such a circuit is called a *multivibrator*.

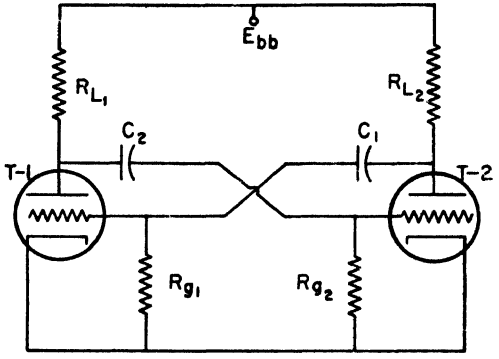


Fig. 2.17. Free-wheeling plate-coupled multivibrator.

The characteristic curve for a voltage-controlled negative resistance, as shown in Fig. 2.18, indicates, as discussed in Art. 2.1, that triggering may be accomplished by changing any of the electrode potentials, such as the grid voltage. In order for free wheeling or

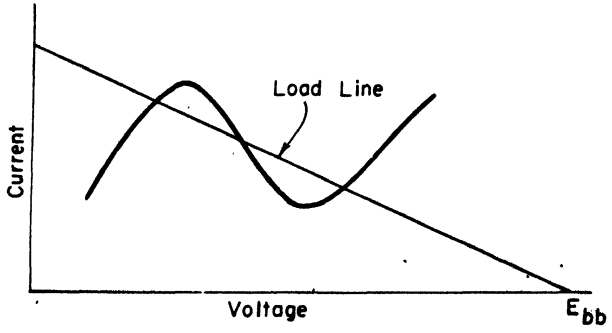


Fig. 2.18. Voltage-controlled negative resistance characteristic.

self-triggering to occur, some electrode potential, such as the grid voltage, must vary over the range embracing the two points where instability occurs, the variation being produced *within* the trigger circuit. Removal of the d-c coupling from the Eccles-Jordan circuit makes this possible.

The method by which the self-triggering action occurs is probably best understood from an analysis of the circuit when the power supply voltages are first applied. Assuming this situation, currents flow in both plate circuits and charges build up on the coupling condensers C_1 and C_2 as the plate voltages rise. If the two halves of the multivibrator were absolutely identical, the currents and charges would build up at exactly the same rate and in the same amount. If, however, due to some slight unbalance arising from thermal agitation, or any other source, a slight increase in current through either tube occurs, instability results.

Assume both tubes to be conducting and that an increase in plate current takes place in tube T-2. Then, the usual cumulative switching action of the trigger circuit takes place, rendering T-2 highly conducting and T-1 cut off. If the initial increase in current had occurred in T-1, the reverse condition of stable equilibrium would have been obtained.

The previously accumulated charge on C_1 now leaks off through R_{g1} so that the grid voltage increases exponentially toward zero. Eventually the grid voltage reaches the tube cutoff voltage and T-1 conducts. When this occurs, the cumulative switching action again occurs, cutting T-2 off and rendering T-1 highly conducting. Then a similar process takes place with C_2 as it discharges until the circuit triggers off again. Hence, the circuit is oscillating at its own natural frequency.

In essence, then, the multivibrator consists of two tubes coupled by an *R-C* circuit. Each tube represents a resistance-coupled amplifier, and since the output of each tube is connected to the input of the other tube, a large amount of positive feedback is used. Generally speaking, two types of coupling are used:

(1) *Plate coupling*—the plate of each tube is connected through the coupling circuit to the grid of the other tube.
(2) *Cathode coupling—one* of the plate-to-grid couplings is changed to cathode-to-cathode coupling.

2.8 Calculation of Multivibrator Waveforms

The analysis of the multivibrator proceeds in the usual manner, that is, by replacing the tubes by switches and equivalent resistances and then reducing the overall equivalent circuit down to the equiva-

lent charging and discharging circuits for each coupling condenser. Proceeding according to this method yields the circuit diagrams of Figs. 2.19 and 2.20.

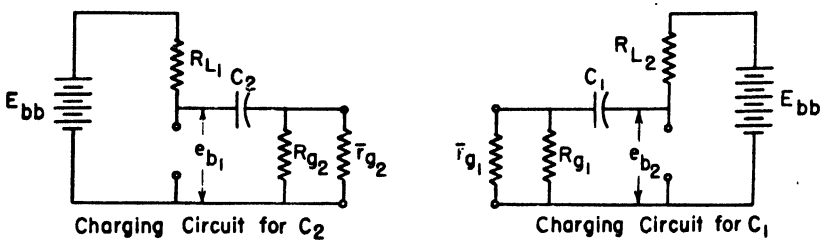


Fig. 2.19.

Consider Fig. 2.19, which is the charging circuit for C_2. During the previous period when C_2 was discharging, the voltage on the grid side of C_2 is $(-E_{c_0})$ since T-2 is cut off, but just getting ready to fire, at the instant considered. The voltage on the plate side of C_2 is the quiescent plate voltage e_{b_0} where

$$e_{b_0} = E_{bb}\left(\frac{\bar{r}_{p_1}}{\bar{r}_{p_1} + R_{L_1}}\right) \tag{2.5}$$

e_{b_0} could be obtained more exactly from the tube characteristics by assuming that the grid voltage on T-1 is 0. Consequently, the voltage across C_2 just at the instant T-2 fires is

$$\gamma_2 = e_{b_0} + E_{c0} = E_{bb}\left(\frac{\bar{r}_{p_1}}{\bar{r}_{p_1} + R_{L_1}}\right) + E_{c0} \tag{2.6}$$

This is the initial condenser voltage for the circuit of Fig. 2.19. Hence, the grid voltage e_{c_2} is

$$e_{c_2} = i_{ch}R_e = \left(\frac{E_{bb} - \gamma_2}{R_e + R_{L_1}}\right)R_e\, \epsilon^{-t/[(R_e+R_L)C_2]} \tag{2.7}$$

where

$$R_e = \left(\frac{\bar{r}_{g_2}R_{g_2}}{\bar{r}_{g_2} + R_{g_2}}\right) \tag{2.8}$$

Hence, at $t = 0$, the initial instant for the charging period of C_2,

$$e_{c_2} = \left(\frac{E_{bb} - \gamma_2}{R_e + R_{L_1}}\right)R_e = (\Delta I)_1 R_e \tag{2.9}$$

where

$$(\Delta I)_1 = \left(\frac{E_{bb} - \gamma_2}{R_e + R_{L_1}}\right) \tag{2.10}$$

The plate voltage on T-1 is

$$e_{b_1} = E_{bb} - e_{R_{L_1}} = E_{bb} - i\,R_{L_1} \tag{2.11}$$

or

$$e_{b_1} = E_{bb} - \left(\frac{E_{bb} - \gamma_2}{R_e + R_{L_1}}\right) R_{L_1}\, \epsilon^{-t/[(R_e + R_{L_1})C_2]} \tag{2.12}$$

and at the initial charging instant

$$e_{b_1} = E_{bb} - (\Delta I)_1 R_{L_1} \tag{2.13}$$

So, e_{b_1} rises exponentially from this value toward E_{bb} while e_{c_2} decreases from its initial value toward zero.

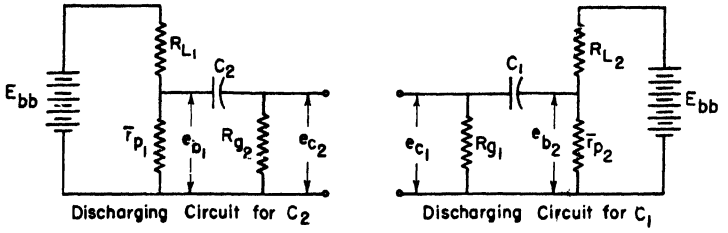


Fig. 2.20.

Now consider the discharging circuit for C_2. Just prior to the instant that T-1 conducts, C_2 had been charged up to a voltage E_{bb} as a result of the preceding charging process so that

$$\gamma_2' = E_{bb} \tag{2.14}$$

where γ_2' = initial condenser voltage for discharging. Figure 2.20 for the discharging of C_2 may be further reduced to that of Fig. 2.21 by application of Thevenin's theorem, where

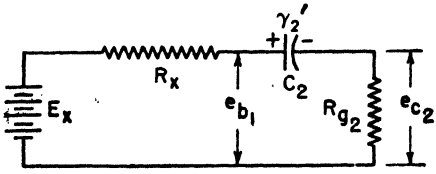


Fig. 2.21.

$$E_x = E_{bb}\left(\frac{\bar{r}_{p_1}}{\bar{r}_{p_1} + R_{L_1}}\right) = e_{b_0} \tag{2.15}$$

and

$$R_x = \left(\frac{\bar{r}_{p_1} R_{L_1}}{\bar{r}_{p_1} + R_{L_1}}\right) \tag{2.16}$$

So, the discharging time constant is

$$T_{\text{dsch}} = C_2(R_x + R_{g_2}) \tag{2.17}$$

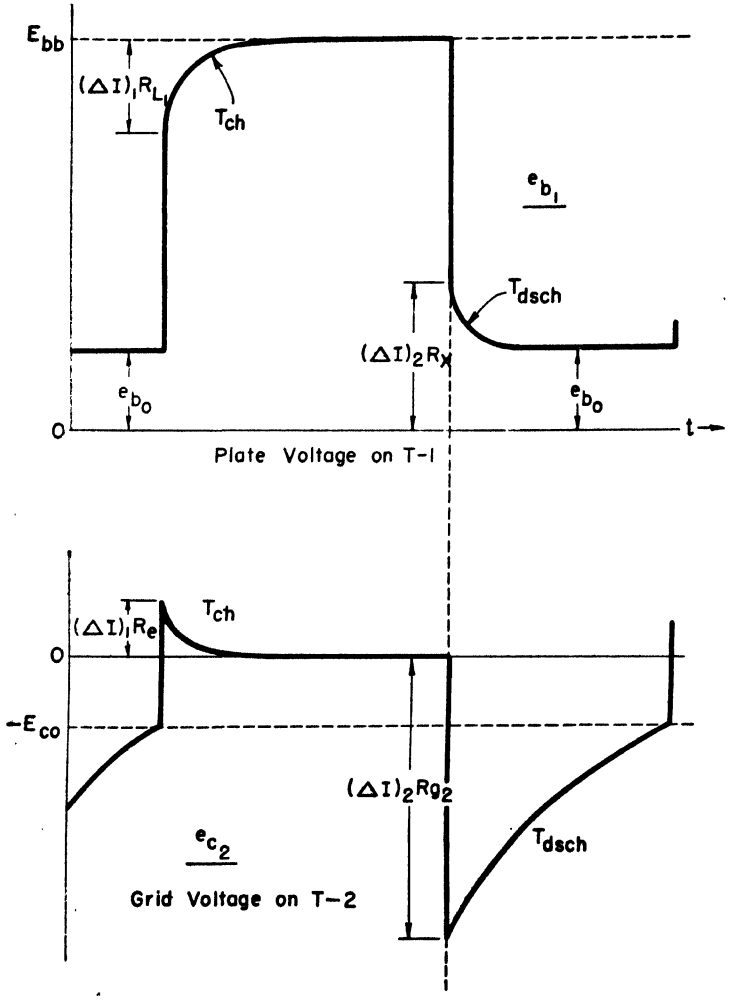


Fig. 2.22(a). Theoretical calculated waveforms for free-wheeling plate-coupled multivibrator.

Consequently, the plate voltage may be obtained as

$$e_{b_1} = E_x - i_{d\,sch} R_x = E_x - (E_x - E_{bb}) \frac{R_x}{R_x + R_{g_2}} \epsilon^{-t/[(R_x + R_{g_2})C_2]} \quad (2.18)$$

$$\text{or} \quad e_{b_1} = e_{b_0} - (e_{b_0} - E_{bb}) \left(\frac{R_x}{R_x + R_{g_2}} \right) \epsilon^{-t/T_{dsch}} \quad (2.19)$$

Now, let $$\left(\frac{e_{b_0} - E_{bb}}{R_x + R_{g_2}}\right) = -(\Delta I)_2 \tag{2.20}$$

Hence, $$e_{b_1} = e_{b_0} + (\Delta I_2)R_x\, \epsilon^{-t/T_{\text{dsch}}} \tag{2.21}$$

Also $$e_{c_2} = i_{\text{dsch}}R_{g_2} \tag{2.22}$$

so $$e_{c_2} = (e_{b_0} - E_{bb})\left(\frac{R_{g_2}}{R_x + R_{g_2}}\right)\epsilon^{-t/T_{\text{dsch}}} \tag{2.23}$$

or $$e_{c_2} = -(\Delta I)_2 R_{g_2}\, \epsilon^{-t/T_{\text{dsch}}} \tag{2.24}$$

And at the initial discharging instant

$$e_{b_1} = e_{b_0} + (\Delta I)_2 R_x \tag{2.25}$$

$$e_{c_2} = -(\Delta I)_2 R_{g_2} \tag{2.26}$$

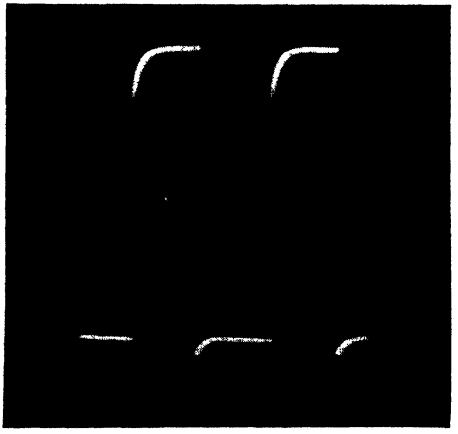
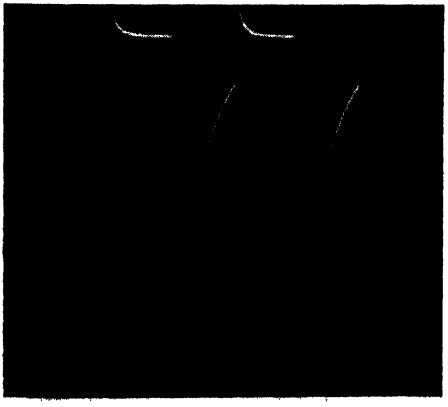


Fig. 2.22(b). Free-wheeling plate-coupled multivibrator plate—voltage waveform.

Fig. 2.22(c). Free-wheeling plate-coupled multivibrator grid—voltage waveform.



Thereafter, the plate voltage on T-1 decreases exponentially toward e_{b_0}, the tube drop, while the grid voltage on T-2 increases exponentially toward zero and the waveforms appear as shown in Fig. 2.22(a).

Similar waveforms and relations could be obtained for the other half of the multivibrator by the same method of analysis.

Two factors have been neglected in the calculations. The first omission is that no account was taken of the grid pip, which is positive, generated by the other tube when it cut off. The positive pip applied to the grid of the tube under consideration drives the tube into heavy conduction, lowering the plate voltage as shown in Fig. 2.23. The quantity $(\Delta I)_2 R_x$ is usually of the order of 2 volts with a time constant of the order of say

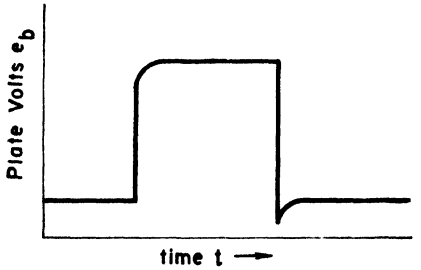


Fig. 2.23. Actual plate-voltage waveform in a free-wheeling plate-coupled multivibrator.

1000 μ sec. On the other hand, the positive pip in the grid circuit is of the order of 5 to 8 volts. Such a signal will produce a negative pip in the plate circuit of sufficient magnitude to obscure the exponential curvature calculated. The resultant waveforms appear as shown in Fig. 2.24.

The other error of omission arose in the assumption that the static plate resistance is a constant. This ties in closely with the discussion in the preceding paragraph because $\bar{r}_p$ actually varies since the grid voltage is initially positive and then decays to zero. This would cause $\bar{r}_p$ to be smaller initially than at the end of the transient and helps to account for the dip in plate voltage discussed above. An exact analysis would require that $\bar{r}_p$ be expressed as a

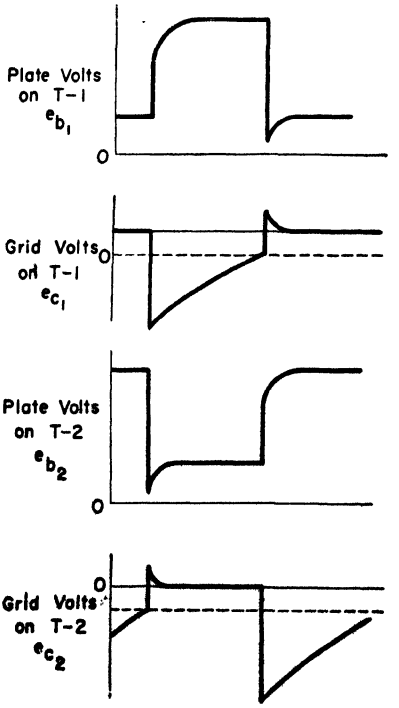


Fig. 2.24. Complete set of waveforms for the free-wheeling plate-coupled multivibrator, showing relative time phase.

function of time as determined by the grid voltage, and then used in the expressions already derived. The increase in accuracy hardly justifies the expenditure of effort, however, as long as the effect can be qualitatively discussed.

2.9 Uncontrolled Frequency of Oscillation of a Multivibrator

The uncontrolled frequency of oscillation of a multivibrator depends primarily upon the time constants of the R-C circuits associated with the grid of each tube, since they determine the length of time between each triggering cycle by holding the appropriate tube cutoff. When the grid of T-1 in Fig. 2.17 is overdriven in the negative region, the grid voltage rises exponentially toward zero as C_1 discharges, reaching cutoff in a length of time roughly proportional to the time constant $R_{g1}C_1$. For T-2, the time between maximum negative grid voltage and cutoff is proportional to $R_{g2}C_2$. Thus, the period of the multivibrator could be expressed implicitly as

$$T = (R_{g1}C_1 + R_{g2}C_2)\,[f(E_{bb})] \tag{2.27}$$

where $f(E_{bb})$ represents the dependence upon the supply voltage through its effects on the tube constants.

The upper frequency limit is imposed by reduction in voltage amplification at the high frequencies. The lower frequency limit arises from, and its value is set by, the leakage resistance in the coupling condensers.

Consider the circuit diagram of a plate-coupled free-wheeling multivibrator as shown in Fig. 2.17. Now, during the time that T-1 is cut off and T-2 is conducting, coupling condenser C_1 discharges through the equivalent circuit shown in Fig. 2.25. The time constant for discharge is

$$T_{\text{dsch}} = C_1(R_{g1} + R_e) \tag{2.28}$$

where $R_e = \left(\dfrac{\bar{r}_{p_2}R_{L_2}}{\bar{r}_{p_2} + R_{L_2}}\right)$ (2.29)

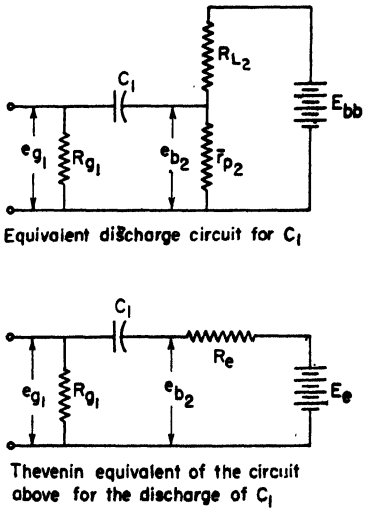


Fig. 2.25.

However, in most application, the static plate resistance is very much less than the grid leak resistance, that is

$$\bar{r}_{p_2} \ll R_{g1}$$

Consequently $$R_e \ll R_{g1}$$

Hence, to a reasonable degree of accuracy,

$$T_{\text{dsch}} = R_{g1}C_1 \tag{2.30}$$

The initial voltage across the condenser is

$$\gamma = E_{bb}$$

Hence, the expression for the grid voltage is

$$e_{g1} = (E_e - E_{bb})\ \epsilon^{-t/R_{g_1}C_1} \tag{2.31}$$

where

$$E_e = E_{bb}\left(\frac{\bar{r}_{p_2}}{\bar{r}_{p_2} + R_{L_2}}\right) \tag{2.32}$$

as obtained from the Thevenized equivalent circuit of Fig. 2.25. So

$$e_{g1} = (E_e - E_{bb})\ \epsilon^{-t/R_{g_1}C_1} \tag{2.33}$$

The discharge ceases when the grid voltage reaches tube cutoff $(-E_{co_1})$. Designating this time as T_1, then

$$-E_{co1} = (E_e - E_{bb})\ \epsilon^{-T_1/R_{g_1}C_1} \tag{2.34}$$

The time T_1 is found as follows:

$$\epsilon^{-T_1/R_{g_1}C_1} = -\frac{E_{co1}}{E_e - E_{bb}} \tag{2.35}$$

$$T_1 = R_{g1}C_1 \ln\left(\frac{E_{bb} - E_e}{E_{co1}}\right) \tag{2.36}$$

Applying the same method to condenser C_2 yields

$$T - T_1 = R_{g2}C_2 \ln\left(\frac{E_{bb} - E_f}{E_{co2}}\right) \tag{2.37}$$

where

$$E_f = E_{bb}\left(\frac{\bar{r}_{p1}}{\bar{r}_{p1} + R_{L_1}}\right) \tag{2.38}$$

and it was assumed that $\bar{r}_{p_1} \ll R_{g_2}$.
Hence, the complete multivibrator period is

$$T = T_1 + (T - T_1) \tag{2.39}$$

or

$$T = R_{g1}C_1 \ln\left(\frac{E_{bb} - E_e}{E_{co1}}\right) + R_{g2}C_2 \ln\left(\frac{E_{bb} - E_f}{E_{co2}}\right) \tag{2.40}$$

Now, if the tubes are identical

$$E_{co_1} = E_{co_2} = E_{co}$$
$$E_e = E_f = E$$

and the expression for the multivibrator period is

$$T = (R_{g_1}C_1 + R_{g_2}C_2) \ln \left(\frac{E_{bb} - E}{E_{co}}\right) \tag{2.41}$$

or

$$T = (R_{g_1}C_1 + R_{g_2}C_2) \ln \left(\frac{E_{bb}}{E_{co}}\right)\left(1 - \frac{E}{E_{bb}}\right) \tag{2.42}$$

This expression is true as long as the grid leak resistances are very much greater than the static plate resistances, and as long as the frequency is low enough so that the effects of the tube interelectrode and wiring capacitances can be neglected. At low frequencies, and under the assumed conditions, accuracies up to 2 per cent can be obtained depending upon the precision of the parts used in construction. The equations may be put in a more simplified form as follows:

since

$$E = E_{bb}\left(\frac{\bar{r}_p}{\bar{r}_p + R_L}\right) \tag{2.43}$$

then

$$\frac{E}{E_{bb}} = \left(\frac{\bar{r}_p}{\bar{r}_p + R_L}\right) \tag{2.44}$$

and hence

$$T = (R_{g_1}C_1 + R_{g_2}C_2) \ln \left(\frac{E_{bb}}{E_{co}}\right)\left(1 - \frac{\bar{r}_p}{\bar{r}_p + R_L}\right) \tag{2.45}$$

or

$$T = (R_{g_1}C_1 + R_{g_2}C_2) \ln \left(\frac{E_{bb}}{E_{co}}\right)\left(\frac{R_L}{\bar{r}_p + R_L}\right) \tag{2.46}$$

Furthermore, within a reasonable engineering degree of accuracy,

$$\frac{E_{bb}}{E_{co}} = \mu \qquad \text{where } \mu \text{ is the amplification factor}$$

so

$$T = (R_{g_1}C_1 + R_{g_2}C_2) \ln \mu \left(\frac{R_L}{\bar{r}_p + R_L}\right) \tag{2.47}$$

Now, when the load resistance is very much greater than the static plate resistance, this expression further simplifies to

$$T = (R_{g_1}C_1 + R_{g_2}C_2) \ln \mu \tag{2.48}$$

This is a fairly accurate rule for predicting the natural frequency and refutes the claim frequently made, that the $f(E_{bb})$ term in the original implicit equation has little effect on the natural frequency. It has, as can be seen, a rather considerable effect.

2.10 Synchronization of Free-Wheeling Multivibrators

Suppose that a small sinusoidal voltage is injected in series in the grid circuit of the free-wheeling plate-coupled multivibrator of Fig. 2.17. If the voltage is inserted in the grid of tube T-1, the net grid voltage on T-1 appears as shown in Fig. 2.26. In the example sketched, the frequency of the synchronizing signal is approximately four times the frequency of the multivibrator since two complete cycles of synch voltage occur during one half cycle of the multivibrator. It is seen that the application of this synch voltage causes the grid to reach cutoff sooner than it would have done so in the absence of the synchronizing signal. Hence, the synch, or control voltage, controls the multivibrator frequency by causing the circuit to trigger off a little sooner than it normally would have. The character of the control voltage is not restricted to sine waves, but can be any periodic waveform conveniently available.

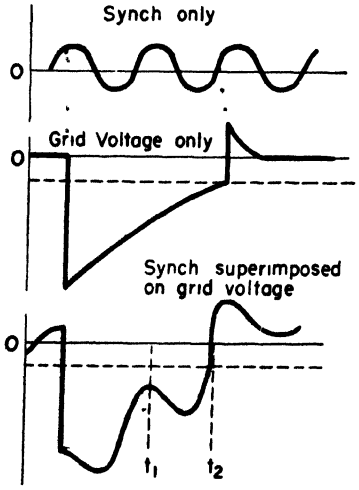


Fig. 2.26. Method of synchronization.

Assuming that the control voltage is applied to only one of the tubes, every complete cycle of the multivibrator will be terminated due to the action of the synch voltage. If, during one complete multivibrator cycle, there are n cycles of the synch voltage, then the controlled frequency of the multivibrator will be f/n where f is the frequency of the synchronizing signal and n, the number of control voltage cycles over one multivibrator cycle.

By inspection of Fig. 2.26, it is evident that if the magnitude of the control voltage were increased sufficiently, the multivibrator could be made to trigger off at t_1 instead of t_2, shortening the multivibrator period by one cycle of the control voltage (remember, synch is being applied to only one grid). Consequently, the controlled frequency of the multivibrator becomes $f/(n-1)$. Hence, for a certain selection of circuit constants, the control voltage can make the multivibrator operate at any one of several submultiples

of the synchronizing frequency, depending upon the relative magnitudes of the injected control voltage.

For a given magnitude of control voltage, there will be a limited range of values of R_g, C, and E_{bb} (which determine the natural frequency of the multivibrator) over which the multivibrator will remain "locked in" with the synchronizing signal. These different values of circuit constants correspond to a range of multivibrator natural frequencies which can be held in synchronism with the control voltage. Increasing the amplitude of the synch voltage will, over a limited range, tend to increase the range of multivibrator natural frequencies which will remain locked in. An experimental determination of this phenomenon characteristically yields curves of the form shown in Fig. 2.27. The two curves indicate the latitude

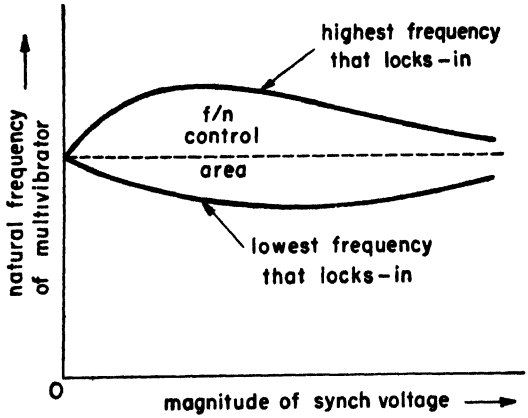


Fig. 2.27. Experimental results using sine-wave synchronization.

of control, and the area represents all possible combinations of R_g, C, and E_{bb}, as well as synch voltage amplitude, which will permit controlled operation. It is frequently referred to as the "f/n control area." Any adjustment which does not fall within this area will result in free wheeling or locking-in at some other multiple of the synch voltage frequency.

The exact form and extent of this control area is governed by a great many factors, but principally

(1) By the degree of symmetry of the multivibrator cycle.

(2) By the method of injecting the synch signal.

Although there are other methods, the two principal methods of injecting synchronization are

(1) Injection to a single grid or its equivalent, injection to both grids with out-of-phase signals.
(2) Injection to both grids with in-phase signals.

The method of synch injection causes the circuit to favor operation at different submultiples of the base frequency.

Before considering the reason for this favoring action it is desirable to examine the mechanism by which control is lost. An example is shown by Fig. 2 28. Curve (1) represents the grid voltage

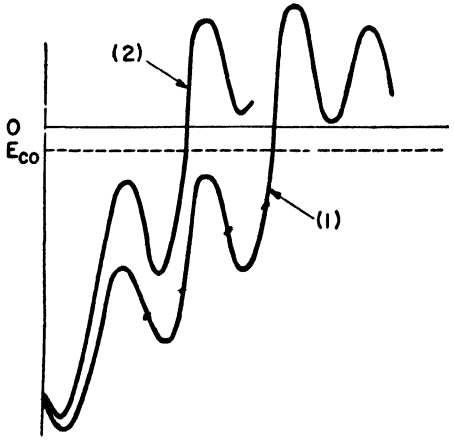


Fig 2 28. Mechanism of loss of control.

of a multivibrator which is locked in at a frequency of one-fifth of the synch voltage frequency. If the natural frequency of the multivibrator is increased, the grid voltage appears as shown by curve (2), indicating that the multivibrator triggers off sooner, near the peak of the synch signal, which results in erratic operation. In other words, operation has passed out of the f/n control area. If the natural frequency were decreased, the same action would occur on the other side of the reference curve.

Now, returning to the favoring action produced by the method of injection of the synch voltage. Assume that a balanced (symmetrical) multivibrator is being synchronized by control voltages

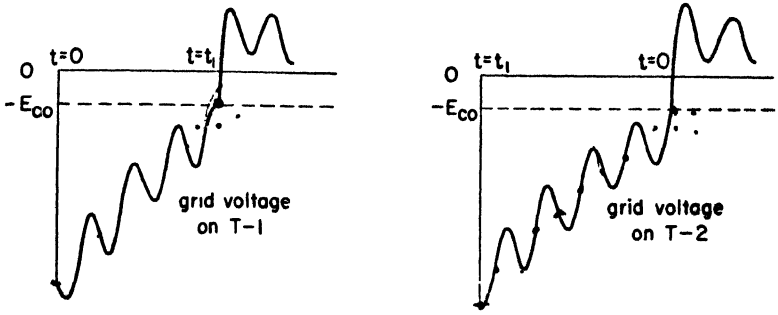


Fig. 2.29. Out-of-phase grid injection at an even ($n = 8$) sub-multiple. Note that T−2 is exerting maximum control, whereas that by T−1 is quite weak (near the peak of the synch voltage).

inserted in series with both grids, but out of time phase. As control is lost by one tube due to an increase in multivibrator natural frequency, it is regained by the other tube at the next submultiple of the control voltage frequency. This means that the multivibrator period is shortened by one cycle of the synch voltage. Hence, it tends to favor operation at the odd submultiples of the base frequency. The action is more apparent from a close examination of Figs. 2.29 and 2.30.

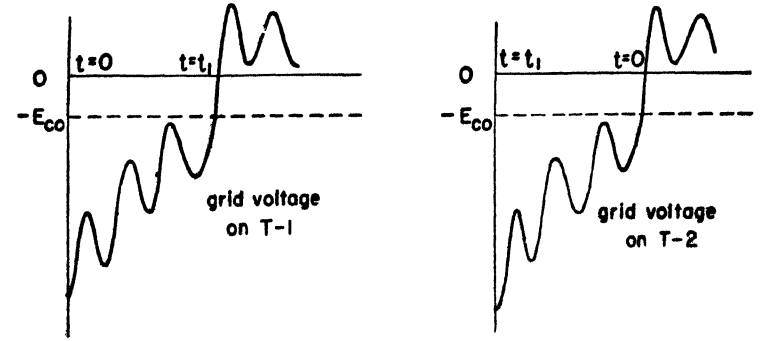


Fig. 2.30. Out-of-phase grid injection at an odd sub-multiple ($n = 7$). Note strong control by both tubes.

The reason for the favoring action may be somewhat more apparent now from the diagrams. Note that in Fig. 2.29, where synchronization is occurring at an even multiple of the control

voltage frequency, and with out-of-phase injection, that only one tube, T-2, is exerting primary control since triggering is occurring on the steep portion of the grid signal. The control by T-1 is rather poor since triggering is occurring near the peak of the synch voltage, a region of unreliability. Thus, as the natural frequency is increased, T-1 picks up stronger control whereas T-2 loses control on the original peak and picks it up on the next. As a result, one cycle of synch is lost. However, now it will be observed (see Fig. 2.30) that both grids are exerting equally strong synchronizing action. Consequently, control is much stronger at the odd submultiples, thus accounting for the favoring action.

If the grids are fed in phase, when locked in, the grid voltage waveforms appear as shown in Fig. 2.31(a) for an even multiple of the

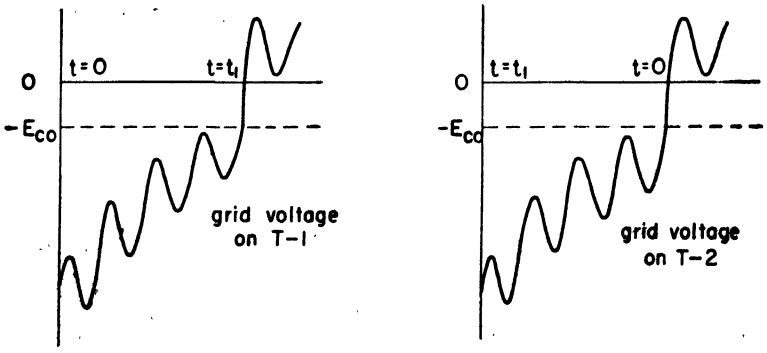


Fig. 2.31(a). In-phase injection at an even sub-multiple ($n = 8$). Note strong control by both tubes.

base frequency. Note that both grids are exerting equal, strong, primary control. An increase in natural frequency would cause both tubes to lose control, and control would *tend* to be picked up by both tubes by each grid shortening its operating cycle by one cycle of synch voltage. Consequently, operation at the even submultiples is favored.

Although the discussion thus far has dealt exclusively with sine wave synchronization, pulses are probably more generally used for this purpose. The space was devoted to sine waves, however, because it is a more difficult idea to grasp, particularly the "favoring action," and because it can frequently be used in lieu of pulses. This

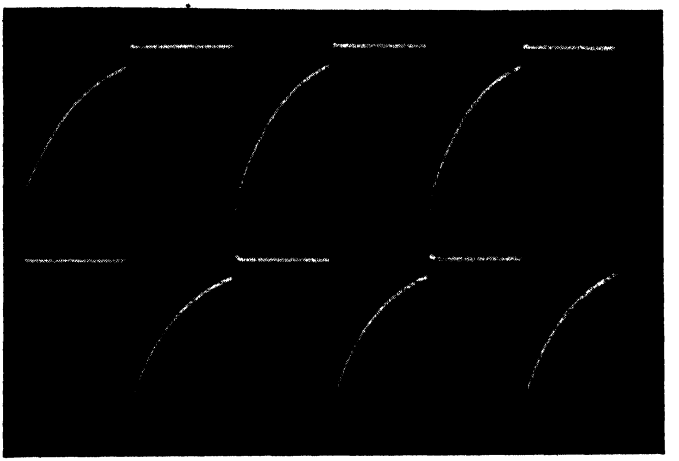


Fig. 2.31(b). Grid-voltage waveforms from a free-wheeling plate-coupled multivibrator, showing relative time phase. No synch voltage applied.

Fig. 2.31(c). In-phase grid injection. Synch frequency is six times multivibrator frequency. Note the strong control exerted by both tubes.

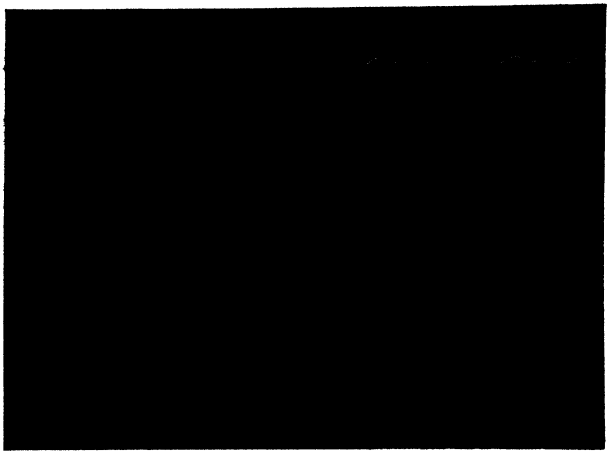


Fig. 2.31(d). In-phase grid injection; $n = 4$. Note that control is about to be lost by both tubes since the point of synchronization is occurring at the synch voltage peaks. An increase in synch voltage amplitude would provide firmer control.

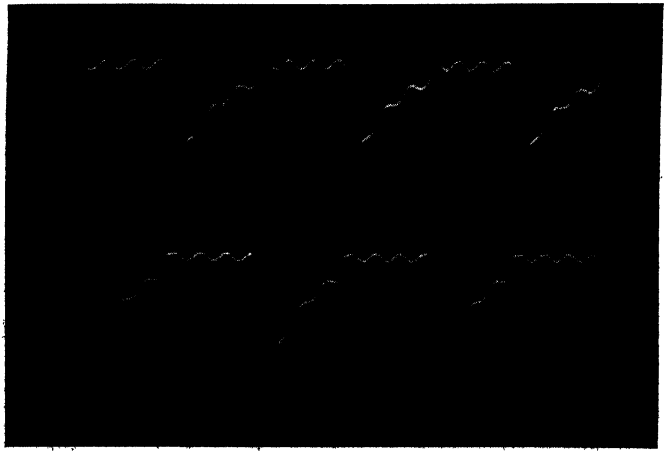


Fig. 2.31(e). In-phase grid injection; $n = 7$. Note that strong control is exercised by only one tube (lower waveform). By comparison between this and figure 2.31(b), it is seen that in-phase injection favors even multiples of the multivibrator frequency.

latter possibility may, in some cases, eliminate tubes that would ordinarily have been required for reshaping an original sine wave into pulses suitable for synchronizing purposes. Generally, where precision is a requirement, pulses are used. A typical waveform obtained with pulse synchronization is shown in Fig. 2.32, which is judged to be self-explanatory.

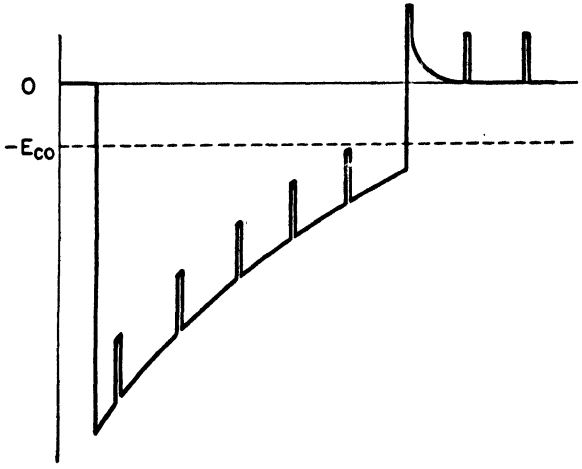


Fig. 2.32. Synchronizing with pulses.

2.11 Effect of Grid Return

The multivibrator circuit discussed so far has had the grid returned to ground through the associated grid leak resistor. This is not, by any means, the only grid return connection possible. Practically, for free-wheeling operation, the grids could be returned to any voltage ranging in value between tube cutoff and the plate supply voltage. By examination of the grid voltage waveforms it is evident that the voltage to which the grid is returned merely determines the rate of approach of the grid voltage toward the cutoff voltage. This is illustrated in Fig. 2.33.

Returning the grid to a higher positive voltage produces two principal effects:

(1) It increases the frequency stability of the multivibrator by increasing the angle at which the grid voltage intercepts the tube cutoff potential.

(2) It shortens the time that the tube remains cut off, assuming the circuit constants to be unchanged.

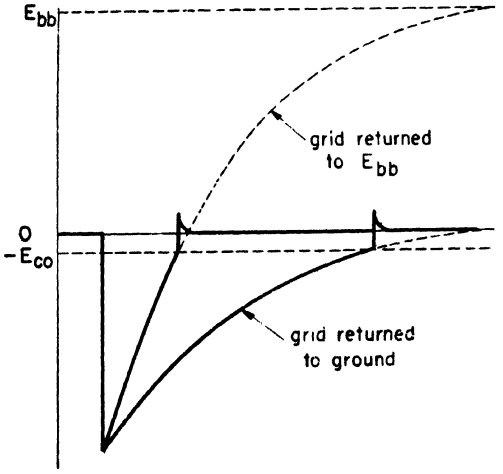


Fig. 2.33.

This latter effect can be used for two purposes:

(1) To develop highly unbalanced outputs, where "unbalance" means that the waveform has a longer positive gate than negative gate, or vice versa. A typical connection is shown in Fig. 2.34 where one grid is returned to ground while the other is returned to the plate supply voltage.

(2) For frequency control by returning both grids to the same variable voltage as indicated in Fig. 2.35.

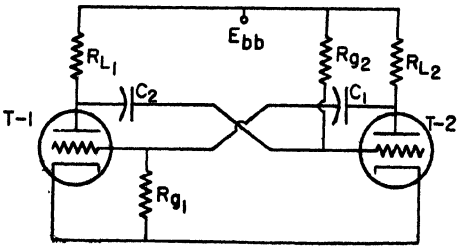


Fig. 2.34. Free-wheeling plate-coupled multivibrator with one grid returned to E_{bb} and the other grid returned to ground.

The first effect, increased frequency stability, can be used in many cases to offset some of the effects of excessive distributed capacitance.

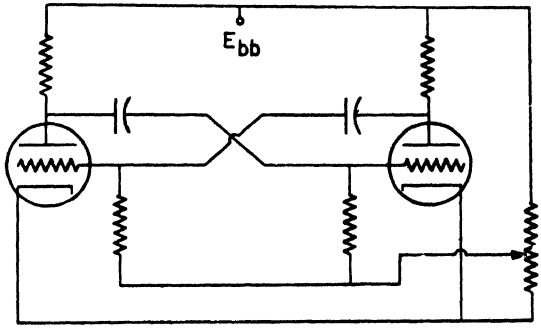


Fig. 2.35. Use of positive-grid return for frequency control of a plate-coupled multivibrator.

2.12 Driven Plate-Coupled Multivibrators

The characteristics of the free-wheeling multivibrator discussed so far are such that it produces self-triggering which arises from a proper combination of electrode potentials. It is possible, however, by adjustment of the polarizing voltages, by any of a variety of schemes, to cause the multivibrator to have one stable point of operation from which it *cannot* trigger itself, and one stable point of operation from which it *can* trigger itself. Consequently, it is normally in the first stable condition and requires some sort of externally applied trigger pulse to initiate each multivibrator cycle. After triggering the action is conventional. Hence, such circuits are usually called "driven multivibrators." Two types will be discussed in this and the next article. The important point to note is the manner in which the polarizing voltages are adjusted to produce the condition outlined above, which condition can exist only if one tube has its grid returned to a voltage well below cutoff, or by development of a suitable cathode voltage.

As a matter of passing interest it should be stated that multivibrators functioning according to the principle above are designated by any of a variety of names, such as *one-shot*, *single-shot*, *single-stroke*, *one-kick*, *flip-flop*, and *start-stop*.

The circuit diagram of a typical driven plate-coupled multivibrator is shown in Fig. 2.36. Single-shot operation is obtained by re-

turning the grid of one tube (T-2 in this case) to a voltage $-E_{cc}$ *which is more negative than the cutoff voltage. Consequently, T-2 is maintained normally off, while T-1 is normally on.*

If a narrow negative pulse is applied to the grid of the normally on tube (T-1), a positive pulse appears at the plate and is coupled over to the grid of the normally off tube (T-2). If this pulse is large enough to bring the grid voltage of T-2 above cutoff, even by a very slight amount, the condition of instability is produced (both tubes

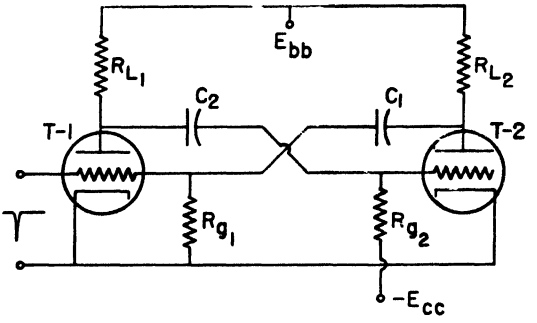


Fig. 2.36. Single-shot plate-coupled multivibrator.

conducting) and the usual cumulative switching action occurs. This renders T-2 conducting and T-1 cutoff.

The coupling condenser C_1 now starts to discharge and T-1 remains nonconducting until enough charge has leaked off the condenser for the grid voltage to reach cutoff, causing current to start in T-1 again. The switching action again occurs, making T-1 conduct and cutting T-2 off.

The charge which accumulated on C_2 while T-2 was conducting now begins to leak off. However, the grid voltage on T-2 never rises as high as cutoff since $E_{cc} > E_{co}$. Consequently, the multivibrator does not self-trigger, but must await another trigger impulse from the external source. Note that this is just another special case illustrating the effect of grid return.

The circuit can be triggered by either of two superficially different methods:

(1) A positive pulse on the grid of the normally off tube.
(2) A negative pulse on the grid of the normally on tube.

However, in both cases, in the final analysis, it is the positive pulse

on the grid of the off tube that causes the triggering since instability can occur only when both tubes conduct at the same time. It is not necessary to cut off the conducting tube with the applied trigger pulse; it simply has to be large enough so that when amplified by the tube it produces a large enough positive pulse on the grid of the nonconducting tube to bring the grid voltage up above cutoff.

The analysis of the circuit and the calculation of waveform proceed exactly as for the free-wheeling case, merely taking cognizance of the fact that the initiation of the multivibrator cycle is due to an external source.

2.13 The Cathode-Coupled Multivibrator

The circuit diagram of a *cathode-coupled multivibrator* is shown in Fig. 2.37. The coupling circuit from the plate of T-1 to the grid of T-2 is conventional except that the grid is returned to the plate supply voltage E_{bb}. However, the usual plate coupling from T-2 to T-1 is replaced by coupling through the mutual cathode resistance R_k. Because of the voltage drop generated in this resistance, voltages measured with respect to ground are not the same as those measured with respect to the cathode.

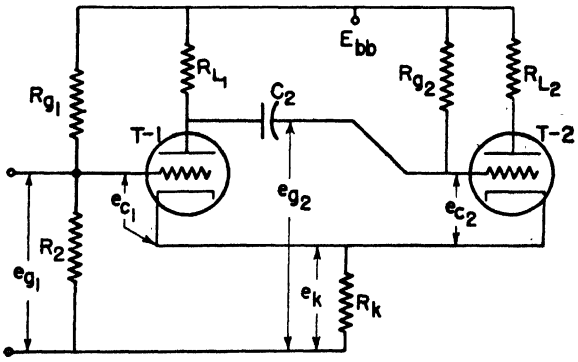


Fig. 2.37. Cathode-coupled multivibrator.

T-2 is maintained normally conducting since its grid is returned to the plate supply. This causes T-2 to conduct heavily, producing a large voltage drop across the common cathode resistance R_k. This makes the cathode positive with respect to ground. Due to the voltage divider action of R_{g_1} and R_2, the grid voltage on T-1 is held

only slightly positive with respect to ground. As a result, the grid-to-cathode voltage on T-1 is negative. The circuit constants are adjusted so that this negative voltage is greater than cutoff, rendering T-1 normally off. Consequently, the stable condition of operation, from which the multivibrator cannot trigger itself, occurs when T-1 is off and T-2 is conducting.

If a positive trigger pulse is applied to the grid of T-1, with sufficient amplitude to cause the tube to conduct, instability occurs and the cumulative switching action results. When current starts to flow in the normally off tube (T-1), a negative pulse appears at the plate and is coupled over to the grid of the normally on tube (T-2). This negative pulse reduces the plate current through T-2, thus reducing the cathode voltage produced by the voltage drop in R_k. This reduction in e_k causes the grid-to-cathode voltage on T-1 (e_{c_1}) to be more positive. This causes more current to flow through R_{L_1}, increasing the amplitude of the negative pulse on the grid of T-2, and so on until the multivibrator reaches its other condition of stable equilibrium.

Since the plate current in T-1 and T-2 both flow through the cathode resistor, the discussion above would appear to be open to question since it does not account for the increase in plate current in T-1 accompanying the decrease in T-2. Actually, the increase in current through T-1 does not offset the decrease in current through T-2 because the grid-to-ground voltage on T-1 is a constant, whereas that on T-2 is varying, decreasing exponentially. As a result, the total current through R_k *does* decrease, causing the cathode voltage to decrease.

T-2 remains cut off until the coupling condenser C_2 can discharge up to the point where conduction of T-2 reoccurs. The switching action then reverses by making e_k increase, which causes e_{c_1} to become more negative, eventually cutting T-1 off.

The particular advantage of this circuit is that the duration of the output pulse may be controlled by adjusting the bias on the grid of T-1. A decrease in grid voltage decreases the width of the output pulse. By proper selection of circuit constants, the output gate width can be made almost a linear function of the voltage e_{g_1}.

For proper operation, e_{g_1} must remain within certain limits:

(1) As e_{g_1} is increased, a point is reached where T-1 cannot be held cut off between trigger pulses, and free-wheeling occurs.

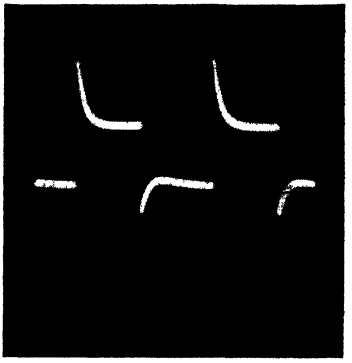


Fig. 2.38(a). Cathode-coupled multivibrator, cathode-voltage waveform.

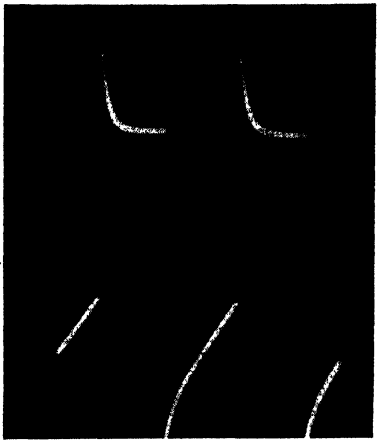


Fig. 2.38(b). Cathode-coupled multivibrator, grid-to-ground voltage (E_{g2}).

(2) As e_{g1} is decreased, the values of e_{g2} at the beginning and end of the pulse approach each other. At the value of e_{g1} for which zero pulse width would nominally occur, the two values of e_{g2} coincide and the multivibrator cannot be triggered.

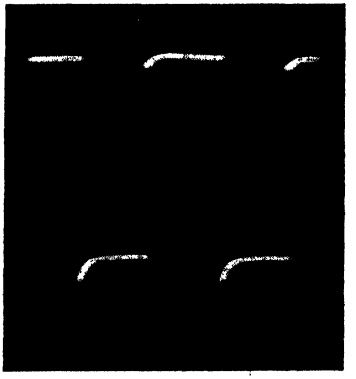


Fig. 2.38(c). Cathode-coupled multivibrator, plate-to-ground voltage (E_{p1}).

Although the calculation of waveform and the analysis of the circuit are somewhat more complex, and certainly trickier in some respects, the method is the same as previously used and does not warrant separate coverage. Typical waveforms are shown in Figs. (2.38) and (2.39).

NOTE: In drawing the load lines for the tubes, remember that the total load includes the cathode resistance.

Fig. 2.39. Cathode-coupled multivibrator, plate-to-ground voltage (E_{p1}). Shows the total range of gate width available by adjustment of the voltage E_{g1}.

2.14 The Blocking Oscillator

Another extremely useful and widely applied trigger circuit is the blocking oscillator. It bears a striking resemblance to the multivibrator in theoretical principle, but the surface dissimilarities appear very great indeed. It will be recalled that the multivibrator consists essentially of two resistance-coupled amplifiers, the output of each feeding the input of the other. The blocking oscillator is similar in that there are two devices, connected in the same manner as in the multivibrator, but one of the devices is a pulse transformer, while the other is a vacuum tube. The pulse transformer is comparable to the second tube of the multivibrator since it provides the phase reversal required for operation. This results in two stable points of operation separated by a region of instability. The operation is such that the tube is overdriven, which causes squaring of the output signal. Consequently, the characteristics of the blocking oscillator are such that it generates narrow pulses. It may be either free-running or driven. It can be used to provide much of the same service as that obtained from a multivibrator, but it has other characteristics which render it more suitable to applications for which the multivibrator is eminently unqualified. For example, blocking oscillators can effectively produce pulses as nrarow as 0.1 μ sec. Such performance from a multivibrator would be little short of miraculous.

The circuit diagram of a typical blocking oscillator is shown in Fig. 2.40. The blocking oscillator tube corresponds to the normally off tube in the driven multivibrator. The pulse transformer acts like the normally on tube.

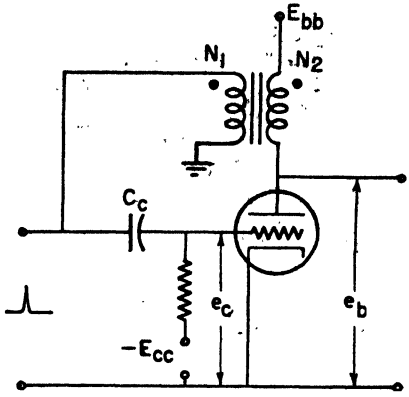


Fig. 2.40. Blocking oscillator.

The blocking oscillator tube is maintained normally off by the bias voltage $-E_{cc}$ on the grid. Before a trigger pulse is applied, no voltage exists across the transformer because the plate current is zero and unchanging. Hence

$$e_b = E_{bb}$$

The coupling condenser (C_c) is charged to the bias voltage ($-E_{cc}$) since one side is connected to the bias source whereas the other side is grounded through the pulse transformer secondary. Hence,

$$e_c = -E_{cc} \qquad \text{grid voltage}$$

Now, if a trigger pulse is applied to the grid of the blocking oscillator tube, of sufficient magnitude to drive the grid voltage above cutoff, plate current commences to flow. Due to the change in current through the transformer primary, a voltage drop is developed, causing the plate voltage to decrease. Since the transformer is connected to give phase reversal, the drop in plate voltage causes the grid voltage to rise, thus causing the plate voltage to decrease even more, and so on. The process is cumulative and in a very short time the tube conducts a large current and the trigger pulse is no longer required to keep the tube on.

The direct equivalent circuit during this period is shown in Fig. 2.41 in which all circuit elements are referred to the plate winding of the pulse transformer. Furthermore, the leakage inductance and distributed shunt capacitance have been neglected. Note the manner of inclusion of the static grid resistance.

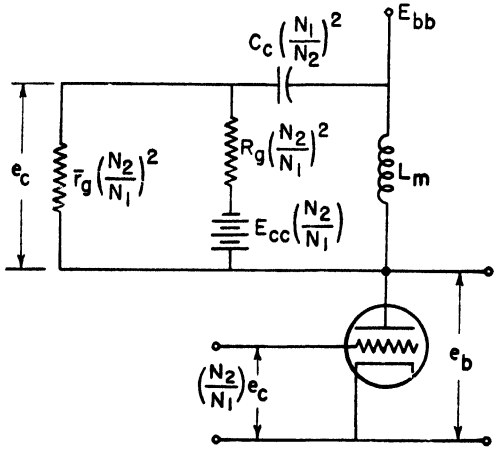


Fig. 2.41. Equivalent circuit at blocking oscillator with all elements referred to the plate winding at the pulse transformer. Leakage inductance and shunt capacitance neglected.

In general, $\bar{r}_g \ll R_g$, so that to a reasonable degree of accuracy, the circuit of Fig. 2.41 could be replaced by that of Fig. 2.42. The initial condenser voltage, designated by γ, is

$$\gamma = \left(\frac{N_2}{N_1}\right)E_{cc} \tag{2.49}$$

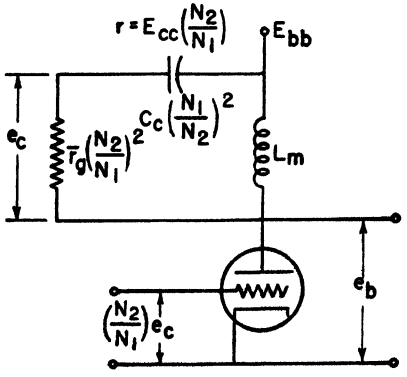


Fig. 2.42. Approximate equivalent circuit of blocking oscillator derived from Figure 2.41 by assuming $R_g \gg \bar{r}_g$.

During the instant immediately after the grid voltage exceeds cutoff, the current through the inductance L_m must still be zero, since the current through an inductance cannot change instantaneously. Consequently, the voltage across the coupling condenser, for the moment, remains $-E_{cc}$. The load line can now be drawn on the tube characteristics for a supply voltage of

$$E_{bb_1} = E_{bb} - \left(\frac{N_2}{N_1}\right)E_{cc} \tag{2.50}$$

and an equivalent load resistance of

$$R_L = \bar{r}_g\left(\frac{N_2}{N_1}\right)^2 \tag{2.51}$$

Now, as the current through the mutual inductance L_m starts to build up, Eq. (2.50) above can be written as

$$E_{bb_1} = e_b + \left(\frac{N_2}{N_1}\right)e_c \tag{2.52}$$

Since E_{bb_1} is a known constant, by assuming a series of values for the grid voltage e_c, the plate voltage can be calculated. A plot can then be made of plate voltage (e_b) vs. grid voltage (e_c) directly on the tube characteristics. Where this graph intersects the load line is the point where the *switching-on* process for the tube ceases. Read the value of e_b at this point and designate it as e_b(min). Then calculate the grid voltage from Eq. (2.52) as

$$e_c(\max) = \frac{N_1}{N_2}\left(E_{bb_1} - e_b \min\right) \tag{2.53}$$

As time passes, C_c charges since current is being supplied through the coupling condenser to the grid of the tube from the transformer. This causes the grid voltage to decrease. Furthermore, the voltage drop across L_m causes the magnetizing current to increase, which in turn causes the plate voltage to drop and the plate current to increase. These two effects cause the point of operation to shift, but at a rate which is slow compared to the previous switching time.

Since both effects occur simultaneously, the analysis is inhibited. However, it can be assumed that the capacitance of the coupling condenser is so large that the voltage across it is practically constant. Thus, the effect of the changing magnetizing current can be isolated

for the purposes of discussion. The approximation is not justified too well in practical circuits, but it offers the only opening to a comparatively straightforward computation of waveform.

As the magnetizing current increases, the load line moves up from the original load line, but remains parallel since the load resistance is constant. This is illustrated in Fig. 2.43 for several assumed values

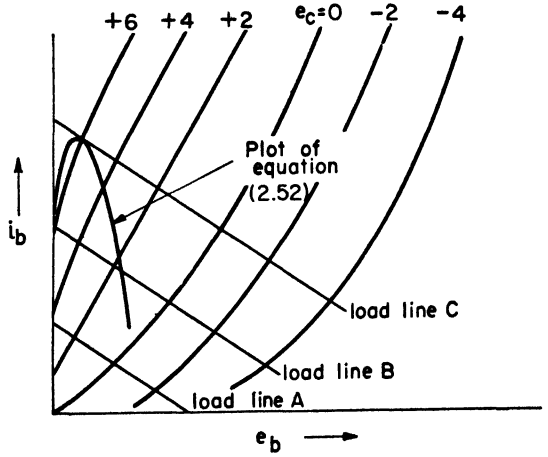


Fig. 2.43. Blocking oscillator analysis.

of current. For any increment in current a new load line can be drawn, the plate voltage found and the grid voltage calculated from Eq. (2.52).

Eventually the increase in current moves the point of operation up to a point where the load line just barely touches the peak of the curve obtained from the equation

$$E_{bb_1} = e_b + \left(\frac{N_2}{N_1}\right)e_c \tag{2.52}$$

At this point, a further increase in magnetizing current cannot be offset by an increase in plate current. Consequently, this marks the end of the pulse. So, as the plate voltage increases, the plate current now begins to decrease, decreasing the transformer voltage, dropping the grid voltage, which causes the plate voltage to increase even more, and so on. Thus, the regenerative switching-*off* action occurs.

During the pulse, the plate and grid voltages are almost constant because of the shape of the curve obtained from equation (2.52). Actually, e_b increases slightly while e_c decreases. This is due to the fact that the assumption regarding the size of the coupling condenser is not exactly met.

Designating the value of plate voltage at the last operating point just before the switching-off action took place as e_{b_2}, then the average voltage across the mutual inductance during the pulse is approximately

$$e_{Lm} = E_{bb} - e_{b_2} \tag{2.54}$$

but

$$e_{Lm} = L_m \frac{di_b}{dt} \tag{2.55}$$

Designating the value of plate current corresponding to e_{b_2} as i_{b_2}, then the pulse duration is approximately

$$\delta = \frac{i_{b_2}}{(di_b/dt)} \tag{2.56}$$

but

$$\frac{di_b}{dt} = \frac{e_{Lm}}{L_m} = \frac{E_{bb} - e_{b_2}}{L_m} \tag{2.57}$$

so

$$\delta = \frac{(i_{b_2})L_m}{E_{bb} - e_{b_2}} \tag{2.58}$$

This approximation holds accurately only so long as the voltage across the coupling condenser is practically constant during the pulse.

After the tube switches off, the shunt capacitance might form a resonant circuit, with the leakage inductance causing low-amplitude high-frequency oscillations. Oscillations with a large amplitude and at a comparatively low frequency can result from the tuned circuit formed by the shunt capacitance and the mutual inductance. Consequently, the resulting waveforms appear as shown in Fig. 2.44. The oscillation in the grid circuit is superimposed on the exponential rise toward $-E_{cc}$.

2.15 The Relaxation Oscillator

Any trigger circuit which is capable of self-triggering is a "relaxation oscillator." However, common usage has broken such circuits down into groups such as multivibrators, transitron oscillators, and so on. As a result, and speaking very broadly, the term, "re-

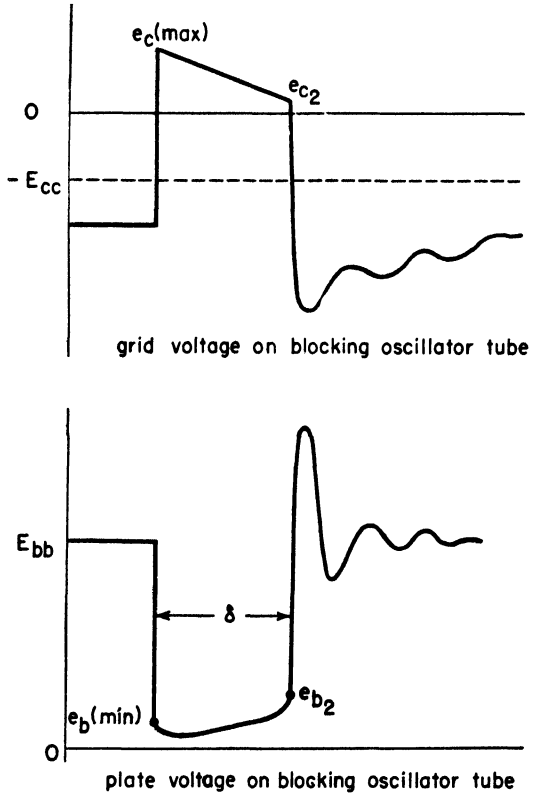


Fig. 2.44. Blocking oscillator waveforms.

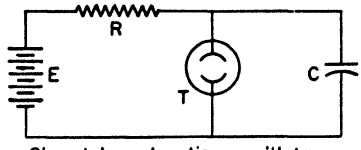
laxation oscillator," when it is not preceded by any other descriptive adjective, has come to signify in many minds the gas tube relaxation oscillator. The term is so applied here.

The "glow-tube relaxation oscillator" has the general circuit diagram shown in Fig. 2.45. The output voltage is developed across the condenser C which is charged by the battery E through the resistor R. T is a gas-filled glow-tube which ignites when the voltage across it exceeds a certain critical value called the "ignition potential" (V_i), and extinguishes when the voltage across it drops to another critical value called the "extinction potential" (V_e). Hence, it acts as a switch, being open when extinguished, and a low resistance when conducting.

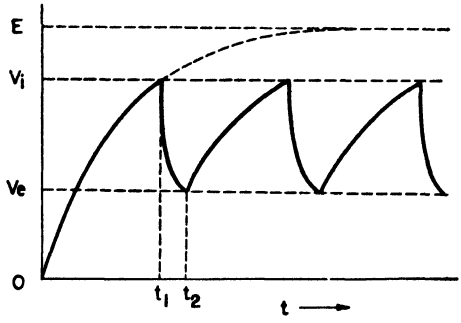
Assume that E is applied to the circuit at the time $t = 0$. Assume that the initial condenser voltage is zero. Hence, the output voltage is

$$e_c = E(1 - \epsilon^{-t/RC}) \tag{2.59}$$

At time $t = t_1$, the output voltage reaches the ignition voltage V_i and the glow-tube fires, throwing a low shunting resistance across the condenser. This causes the condenser to discharge rapidly, reducing the voltage across the tube until the voltage reaches the extinction potential. The tube stops conducting, removing the low shunting resistance, and the condenser charges up again. The cycle repeats indefinitely.



Glow tube relaxation oscillator



Relaxation oscillator output

Fig. 2.45. Relaxation oscillator.

By the method of Art. 2.9, it can be shown that the frequency of oscillation is given by the following expression, assuming the discharge time to be negligible compared to the charging time:

$$f = \frac{1}{RC \ln \left(\frac{E - Ve}{E - Vi}\right)} \tag{2.60}$$

Since the ignition and extinction potentials are normally complicated functions of frequency, tube current, temperature, and electrode illumination, the formula is of little value in *predicting* the frequency of oscillation.

In most practical circuits the glow tube is replaced by a thyratron (such as an 884) so that synchronization may be effected by injection on the control grid. The fundamental considerations are then the same as for the synchronization of the free-wheeling multivibrator.

2.16 A Table of Practical Values for the Circuits of Chapters 1 and 2

Diode Clipper (Fig. 1.10)

E_{bb}	80 volts
Tube	½6H6
R_L	100 K

Cathode Follower (Fig. 1.17)

E_{bb}	250 volts
Tube	½6SN7
R_k	25 K
R_c	500 K

R-C Peaker (Fig. 1.28)

R	2500 ohms
C	200 $\mu\mu$f

Diode Clamper (Fig. 1.44)

E_{cc}	−50 volts
Tube	½6H6
R_c	1 megohm

Saw-tooth Generator (Fig. 1.63)

E_{bb}	250 volts
Tube	½6SN7
R_L	100 K
C	0.5 μf

Bootstrap Saw-tooth Generator (Fig. 1.70)

E_{bb}	250 volts
Tubes	½6H6 and 6SN7
R_L	200 K
R_k	25 K
C_{fb}	0.5 μf
C	500 $\mu\mu$f

Trapezoidal Voltage Generator (Fig. 1.74)

E_{bb}	250 volts
Tube	½6SN7
R_1	100 K
R_2	25 K
C	0.05 μf

R-L-C Peaker (Fig. 1.77)

E_{bb}	250 volts
Tube	½6SN7
L	3 mhy
R	4700 ohms
C_d	distributed C of coil

Ringing Circuit (Fig. 1.80)

E_{bb}	250 volts
Tube	½6SN7
L	12.4 mhy
C	300 $\mu\mu$f

Eccles-Jordan Trigger Ckt. (Fig. 2.13)

Tube	6SN7
R_L	1 megohm
R_1	4 megohm
R_2	4 megohm
R_g	4 megohm
C_1	50 $\mu\mu$f
C_2	50 $\mu\mu$f

Gas Tube Trigger Ckt. (Fig. 2.15)

Tubes	884's
E_{bb}	110 volts
R_g	100 K
R_1, R_2	3 K
R_k	3 K
C_k	0.0015 μf
C	0.0020 μf

Free-Wheeling MV (Fig. 2.17)

E_{bb}	250 volts
Tube	6SN7
R_L	20 K
R_g	500 K
C	800 $\mu\mu$f

Single-Shot MV (Fig. 2.36)

E_{bb} 250 volts

Same circuit constants as the free-wheeling multivibrator, but E_{cc} is −50 volts.

Cathode-Coupled MV (Fig. 2.37)

E_{bb}	250 volts
Tube	6SN7
R_{L1}	20 K
R_{L2}	10 K
R_{g1}	100 K
R_{g2}	200 K plus 2 meg. pot.
R_2	50 K
R_k	10 K
C_2	2000 $\mu\mu$f

Blocking Oscillator (Fig. 2.40)

E_{bb}	250 volts
Tube	½6SN7
C_c	200 $\mu\mu$f
R_g	1 megohm
E_{cc}	−90 volts

⅓ turns ratio for transformer

PROBLEMS

2.1 Calculate the frequency of a symmetrical multivibrator with grids returned to ground, if

E_{bb} = 300 volts
R_L = 20 K
R_g = 1 megohm
C_c = 1000 $\mu\mu$f
E_{co} = - 18 volts
$\bar{r}_g$ = 1000 ohms for $e_c \geq 0$
$\bar{r}_p$ = 10 K for $e_c \geq 0$

2.2 Calculate and plot all significant and critical points on the plate and grid voltage waveforms of the multivibrator of Problem 1.1.

2.3 A multivibrator is used in a timer to produce a positive rectangular gate 107.3 μ sec long. The multivibrator is synchronized by a 400 cps negative pulse applied to the grid of T-1. The positive gate is to begin at the instant that the synch pulse is applied. The following data are given:

	T-1	T-2
E_{bb}	300 volts	300 volts
Grid leak resistor	500 K	1 megohm
Grid leak connection	To E_{bb}	To cathode
R_L	20 K	30 K
u	20	20
$\bar{r}_g$	500 ohms	500 ohms
$\bar{r}_p$	10 K	10 K
r_p	7 K	7 K

(a) Draw the complete circuit diagram. Indicate the terminals to which the synch pulse is applied, and the terminals from which the positive gate is removed.

(b) Calculate the value of the coupling condenser between the grid of T-1 and the plate of T-2.

2.4 Repeat part (b) of Problem 2.3, but assume that both grid leaks are returned to the cathode.

2.5 Why is it more desirable to connect the circuit as done in Problem 2.3, compared to the connection in 2.4?

2.6 Would a 12-volt trigger pulse be sufficient to synchronize the multivibrator of Problem 2.3?

2.7 What restriction is imposed upon the size of the coupling condenser between the plate of T-1 and the grid of T-2 in the circuit of Problem 2.3?

2.8 Assuming the multivibrator has both grids returned to the cathode, calculate the minimum value of the capacitor of Problem 2.7 which permits proper synchronization from the 400-pps synch signal.

2.9 Assuming the condenser of Problem 2.8 to be 1000 $\mu\mu$f instead of the critical value calculated, what is the minimum magnitude of negative synch pulse that can be used to obtain synchronization at 400 cps?

2.10 When both grid leaks of a symmetrical multivibrator are returned to the same negative bias source, it is found experimentally that the circuit will oscillate for all values of negative bias up to 50 or 60 volts. Yet, the cutoff voltage of the tubes may be only −18 volts. Explain the reason for this action.

REFERENCES FOR CHAPTERS 1 AND 2

Books

Brainerd, J. G., Reich, H. J., Koehler, G., and Woodruff, L. F., *Ultra-High-Frequency Techniques*, Van Nostrand, 1942.

Eastman, A. V., *Fundamentals of Vacuum Tubes*, McGraw-Hill, 1941.

M.I.T. Staff, *Principles of Radar*, McGraw-Hill, 1946.

Lewis, W. B., *Electrical Counting*, Cambridge Univ. Press.

Fink, D. G., *Principles of Television Engineering*, McGraw-Hill, 1940.

Puckle, O. S., *Time Bases*, Wiley, 1943.

Terman, F. E., *Measurements in Radio Engineering*, McGraw-Hill, 1935.

Periodicals

Abraham, H., and Bloch, E., *Ann. de Phys.*, **12**, 252 (1919).

Brunetti, Cledo, *Proc. IRE*, **27**, 88 (1939).

Carrara, N., *Alta Frequenza*, **8**, 683 (1939).

Chakravarti, S. P., *Phil. Mag.*, **30**, 294 (1940).

Eccles, W. H., and Jordan, F. W., *Radio Rev.*, **1**, 143 (1919).

Herold, E. W., *Proc. IRE*, **23**, 1201 (1935).

Hull, L. M., and Clapp, J. K., *Proc. IRE*, **17**, 252 (1929).

Hunt, F. V., *Rev. Sci. Instr.*, **6**, 43 (1935).

Jofeh, M. L., *Journall IEE*, **85**, 400 (1939).

le Corbeiller, P., *Journall IEE*, **79**, 361 (1936).

Reich, H. J., *Rev. Sci. Instr.*, **9**, 222 (1938).
Electronics, Aug. (1939).

Van der Pol, B., *Phil. Mag.*, **6**, 978 (1920).
Phil. Mag., **2**, 978 (1926).
Proc. IRE, **22**, 1051 (1924).

Vecchiacchi, F., *Alta Frequenza*, **9**, 745 (1940).

Watanabe, Y., *Proc. IRE*, **18**, 327 (1930).

CHAPTER 3

AMPLIFICATION FOR UHF SYSTEMS

THE amplification of signals incident to proper operation of a radar or television receiver, or similar video device, imposes much more stringent restrictions upon the designer than he is ordinarily subjected to in audio-frequency amplifier construction and design. The differences arise largely from the fact that the ear has very limited discriminatory powers: its normal frequency response seldom exceeds a bandwidth of about 30 cps to 15 kcps. Satisfactory sound reproduction may actually be obtained over a considerably more restricted range, as evidenced by the ordinary telephone receiver. Furthermore, the ear does not distinguish small relative phase shifts. At least, as long as the bandwidth requirements are met, the phase shift is seldom objectionable and usually is not considered in the design. Thus, the problem of aural amplification is largely one of passing the desired band of frequencies.

In dealing with the eye, which is a vastly more demanding mechanism, waveform must be preserved to a very high degree of accuracy, thus necessitating the amplification of a wide band of frequencies at constant time delay. As an example, consider a television picture which, according to present practice, is divided up into 525 horizontal strips and transmitted 30 times per second. According to RMA standards, when the picture is composed of light elements whose dimensions bear the following relationship to one another,

$$\frac{\text{horizontal dimension}}{\text{vertical dimension}} = \frac{4}{3}$$

then the picture detail is a maximum. Hence, there are $525 \times 525 \times \frac{4}{3}$, or 367,500, light and dark patches transmitted every $\frac{1}{30}$th of a second. That is, $30 \times 367{,}500$, or 11,025,000, light and dark spots sent each second. Assuming each light and dark spot to represent one cycle of the fundamental, then the fundamental frequency of the television signal is something in excess of 5 mcps. The lower fre-

quency limit is about 30 cps, the number of pictures sent per second. Thus, a wide-band amplifier is needed, and the transmission must occur with constant time delay in order to eliminate phase distortion.

Under such restrictions, design becomes infinitely more complex. The physical arrangement of components, seldom a major consideration in audio amplifiers, becomes extremely important in video amplifiers, imposing severe limitations on the freedom of the designer. The over-all problem involves amplification under two separate sets of conditions:

(1) Amplification of the video signal alone.
(2) Amplification of the video signal that has modulated an RF or IF carrier.

The straight video amplifier is usually resistance-coupled and frequently employs compensating circuits of various types. Conventionally, IF and RF amplifiers use single- or double-tuned circuits for coupling purposes. In certain cases, the tuned circuits may be resonant lengths of transmission line.

3.1 Fourier Analysis of a Periodic Rectangular Pulse

The input to a radar or television amplifier consists ideally of short rectangular pulses. For a radar set, the pulse duration is normally fixed, but the frequency of the received signal would be somewhat erratic since it would depend upon the number and range of the targets under consideration. In a television amplifier, the signal may be considered to consist of pulses of varying width and frequency. However, it is possible to specify some minimum pulse

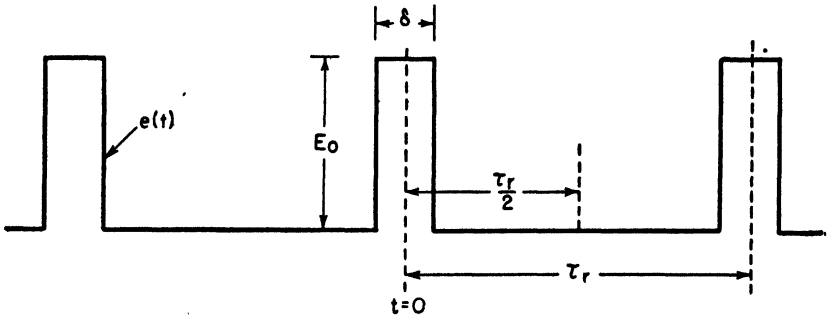


Fig. 3.1. Recurrent pulse input to a video amplifier—specified as a function $E(t)$.

width and frequency of recurrence for both the television and radar video amplifiers which will then set the minimum design requirements. Thus, assuming the minimum pulse width to be δ, and the lowest recurrence rate as f_r, then the input to the amplifier could be conveniently pictured as shown in Fig. 3.1.

For the sake of convenience, select the initial point of observation of the wave ($t = 0$) as shown in the figure. Furthermore, assume that the period between centers of the recurring pulses is much greater than the pulse width; that is,

$$\tau_r \gg \delta$$

This is a good approximation in actual practice since δ is normally of the order of a microsecond, whereas τ_r is of the order of several thousand microseconds.

The voltage $e(t)$ may be written in the form of a Fourier cosine series as

$$e(t) = \sum_{K=0}^{\infty} A_n \cos a_n t \tag{3.1}$$

The series is a cosine series because the initial point of observation was selected in such a way that the voltage function $e(t)$ is symmetrical about the $t = 0$ line, that is, it is an *even* function. For an even function, the unknown terms A_n and a_n can be calculated from the following equations:*

$$a_n = \frac{n\pi}{\frac{1}{2}\tau_r} = \frac{2n\pi}{\tau_r} = 2n\pi f_r \tag{3.2}$$

$$A_n = \frac{2}{\frac{1}{2}\tau_r} \int_0^{\frac{1}{2}\tau_r} e(t) \cos (2\pi n f_r t)\, dt \tag{3.3}$$

Before substituting into the integral, note that the function $e(t)$ has a value of E_0 for a period of time $\delta/2$ following the initial time of observation. Thereafter, $e(t)$ is zero. This may be expressed mathematically as follows:

$$e(t) = E_0 \quad \text{for} \quad 0 < t < \frac{\delta}{2} \tag{3.4}$$

$$e(t) = 0 \quad \text{for} \quad \frac{\delta}{2} < t < \frac{\tau_r}{2} \tag{3.5}$$

*See Churchill, R.V., *Fourier Series and Boundary Value Problems*, p. 74, McGraw-Hill.

Substituting these boundary conditions into the integral of Eq. (3.3) yields

$$A_n = \frac{4}{\tau_r}\int_0^{\delta/2} E_0 \cos(2\pi n f_r t)t + \frac{4}{\tau_r}\int_{\delta/2}^{\tau_r/2} 0 \cdot dt \tag{3.6}$$

However, the second term is the integral of zero, which is itself zero, so that Eq. (3.6) becomes

$$A_n = \frac{4E_0}{\tau_r}\int_0^{\delta/2} \cos(2\pi n f_r t)\, dt \tag{3.7}$$

Integrating yields

$$A_n = \left(\frac{4E_0}{\tau_r}\right)\left(\frac{1}{2\pi n f_r}\right)\left(\sin 2\pi n f_r t\right)_0^{\delta/2} \tag{3.8}$$

When the limits are substituted, the following equation results:

$$A_n = 2E_0\left(\frac{\delta}{\tau_r}\right)\left(\frac{\sin n\pi f_r \delta}{n\pi f_r \delta}\right) \tag{3.9}$$

Substituting the values of a_n and A_n, calculated in Eqs. (3.2) and (3.9), into Eq. (3.3) yields

$$e(t) = 2E_0\frac{\delta}{\tau_r}\sum_{n=0}^{\infty}\left(\frac{\sin n\pi f_r \delta}{n\pi f_r \delta}\right)\cos(2\pi n f_r t) \tag{3.10}$$

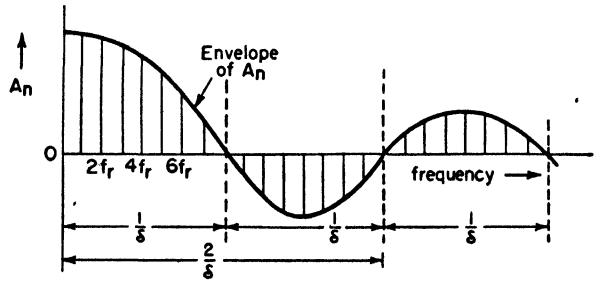


Fig. 3.2. Spectrum amplitude coefficient for the pulse of Fig. 3.1.

The frequency dependence of the amplitude coefficient A_n may be represented as shown in Fig. 3.2. The first component is located at f_r, the second at $2f_r$, the third at $3f_r$, and so on, uniformly spaced

at intervals of f_r. Further examination of the spectrum amplitude coefficient shows that:

(1) Increasing the pulse repetition frequency (*PRF*) decreases the number of components under each loop without altering the spacing of the loops.

(2) Increasing the pulse width decreases the space between loops, and therefore decreases the number of components under the loop without changing the spacing between components.

By integrating the equation for A_n it is possible to show that the area of the first loop *relative* to the second is a constant, independent of the pulse dimensions. The same is true for succeeding loops. Consequently, the frequencies at which the envelope passes through zero determine the relative importance of the frequency components. These frequencies depend only upon the pulse width δ, being $1/\delta, 2/\delta, 3/\delta$, and so on. Hence, the high-frequency response necessary for a given quality of reproduction is determined by the pulse width δ, whereas the lowest frequency to be amplified is governed by the *PRF*. Consequently, the quality of reproduction is not determined by the number of harmonics that are passed, but by the number of loops in the amplitude coefficient that are passed.

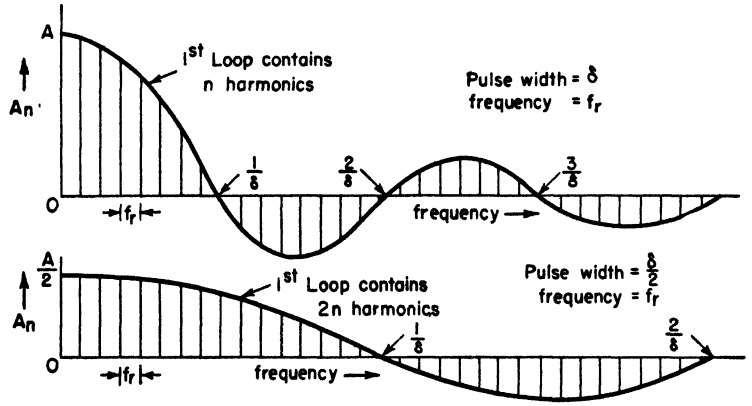


Fig. 3.3. Effect of variation in pulse width on the spectrum amplitude coefficient.

A change in the pulse duration δ produces the effect on the amplitude coefficient shown in Fig. 3.3. It is apparent that changing the pulse duration does not affect the harmonic spacing, but causes more

harmonics to be included under a single loop, with a decrease in the initial magnitude. Hence, a smaller pulse width requires a greater high-frequency response to obtain the same quality of reproduction. In the case shown in Fig. 3.3, the pulses will be passed with equal fidelity if the same number of loops (not harmonics) are transmitted. Thus, halving the pulse width requires doubling the bandwidth. As a result, it can be concluded that the rate of repetition of the pulse, the pulse duration, and the degree of fidelity with which it must be reproduced are the factors which determine the bandpass required for pulse amplification. Furthermore, this amplification must be obtained at constant time delay. Practical amplifiers do not always possess these characteristics due to the inherent limitations associated with them, but a study of the nature and extent of the limitations suggests avenues through which improvements can be made. Consequently, a brief review of ordinary resistance-coupled amplifiers is indicated before proceeding to the more advanced material.

3.2 Review of Distortion

Any four-terminal network can be characterized by means of its input-output characteristic, or the complex voltage ratio, which gives the attenuation and phase shift characteristics of the network as a function of frequency. A typical case was worked out in detail in Art. 1.7 of Chap. 1. Assuming that a rectangular pulse of the form discussed in the preceding article is applied to a four-terminal network, from the Fourier analysis of the pulse it is possible to write down the conditions which the network must fulfill if distortionless transmission is to be obtained.

Consider one harmonic, of frequency f, of this pulse. In passing through the four-terminal transducer it will be attenuated by α nepers and will experience a phase lag of β radians. A phase lag of 2π radians corresponds, in time, to one period T of the harmonic. So, letting t_0 designate the actual time delay, by proportion we can write

$$\frac{t_0}{T} = \frac{\beta}{2\pi} \tag{3.11}$$

Now, solve for the actual time delay t_0, obtaining

$$t_0 = T\left(\frac{\beta}{2\pi}\right) \quad \text{but} \quad T = \frac{1}{f} \tag{3.12}$$

so

$$t_0 = \frac{\beta}{2\pi f} = \frac{\beta}{\omega} \tag{3.13}$$

Now, if the attenuation α and time delay t_0 are both pure constants, independent of frequency, all the harmonics in the input will be attenuated to the same degree and delayed in time by the same amount. As a result, the output waveform will have exactly the same shape as the input, but will be changed in amplitude and displaced t_0 sec in time. Consequently, it is apparent that two conditions must fulfilled for distortionless transmission:

(1) The attenuation α must be independent of frequency.
(2) The time delay t_0 must be independent of frequency, or, β must be a linear function of frequency.

These requirements are illustrated by Fig. 3.4.

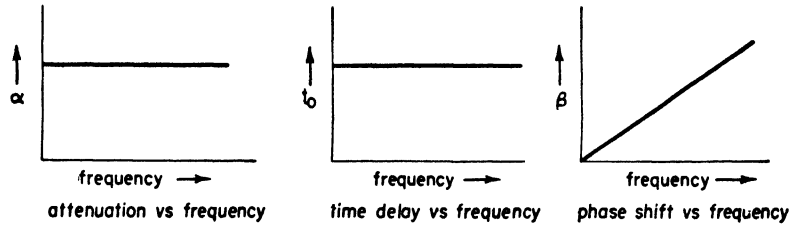


Fig. 3.4. Requirements for no distortion.

Such idealized transmission is, of course, never obtained, so that signals traversing transducers are always distorted from their original shape. The type of distortion introduced depends upon which of the above idealized characteristics is violated. Thus, two types of distortion are distinguished in linear circuits:

(1) Amplitude distortion—when α varies with frequency.
(2) Phase (or delay) distortion—when t_0 varies with frequency.

A knowledge of the origin of the deviations from the ideal in amplifiers is necessary if an intelligent appraisal of the situation is to be made.

3.3 Review of Resistance-coupled Amplifiers

The circuit diagram of a typical resistance-coupled amplifier is shown in Fig. 3.5. Distributed and interelectrode capacitances bearing upon the problem are shown by dotted lines. Application of the equivalent plate circuit theorem yields the equivalent plate circuit of Fig. 3.6 where the tube has been replaced by its Thevenin's theorem equivalent.

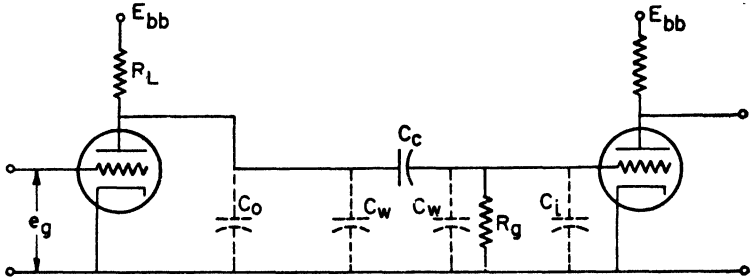


Fig. 3.5. Resistance-coupled voltage amplifier.

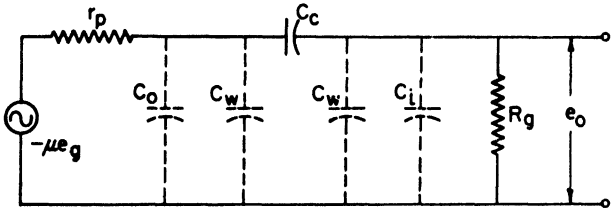


Fig. 3.6. Equivalent plate circuit of Fig. 3.5.

In the usual introductory material on amplifiers it is shown that the voltage gain K is given by

$$K = \frac{e_0}{e_g} = -\frac{g_m Z_L}{1 + (Z_L/r_P)} \tag{3.14}$$

where Z_L is the complex transfer impedance in the plate circuit. Due to the generally complicated nature of this impedance in Fig. 3.6, it is found expedient to divide the over-all problem into three special cases:

(1) High frequency.
(2) Mid-frequency.
(3) Low frequency.

This simplifies the problem considerably. The frequency range in which the reactance of the coupling condenser C_c is appreciable compared to the grid leak resistance R_g is called the *low-frequency* range. The frequency range in which the shunting effect of the distributed and interelectrode capacitances is appreciable is designated as the *high-frequency* range. In this range the reactance of the coupling condenser is practically zero and can be neglected. In the inter-

mediate range between these two extremes, the *mid-frequency* range, the reactance of the coupling condenser is negligibly small compared to the grid leak resistance whereas the susceptance of the shunt capacitances is small compared to the parallel combination of the plate load resistance and the grid leak. This leads directly to the three equivalent circuits of Fig. 3.7.

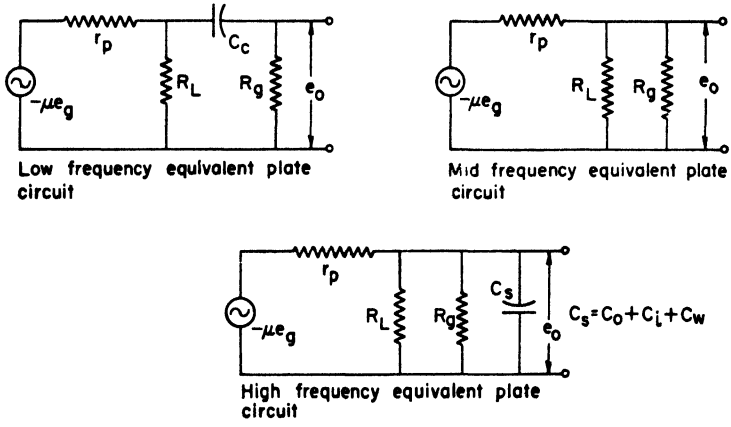


Fig. 3.7. Breakdown of the equivalent plate circuit of Fig. 3.6, as a function of frequency.

Solution of these circuits produces the following expressions for the absolute magnitude of the voltage gain and relative phase angle.

$$X_c = \frac{1}{\omega C_c}$$

$$|K_L| = \frac{|K_m|}{\sqrt{1 + (X_c/R'_{eq})^2}} \qquad R'_{eq} = \frac{r_p R_L + r_p R_g + R_L R_g}{r_p + R_L} \tag{3.15}$$

$$\tan \theta_L = \frac{X_c}{R'_{eq}} \tag{3.16}$$

$$|K_m| = g_m R_{eq} \qquad R_{eq} = \frac{r_p R_L R_g}{r_p R_L + r_p R_g + R_L R_g} \tag{3.17}$$

$$\tan \theta_m = 0 \tag{3.18}$$

$$X'_c = \frac{1}{\omega C_s} \tag{3.19}$$

$$|K_H| = \frac{|K_m|}{\sqrt{1 + (R_{eq}/X'_c)^2}} \qquad C_s = \text{total shunt capacitance}$$

$$\tan \theta_H = -\frac{R_{eq}}{X'_c} \tag{3.20}$$

The frequencies at which the power is one half its mid-frequency value (3 db down), or where the voltage gain is down to 0.707 of its maximum value, are obtained as follows:

(1) The lower half-power frequency (f_1) occurs when

$$R'_{eq} = X_c = \frac{1}{\omega_1 C_c} \tag{3.21}$$

so

$$\omega_1 = \frac{1}{R'_{eq} C_c} \quad \text{or} \quad f_1 = \frac{1}{2\pi R'_{eq} C_c} \tag{3.22}$$

(2) The upper half-power frequency (f_2) occurs when

$$R_{eq} = X'_c = \frac{1}{\omega_2 C_s} \tag{3.23}$$

so

$$\omega_2 \frac{1}{R_{eq} C_s} \quad \text{or} \quad f_2 = \frac{1}{2\pi R_{eq} C_s} \tag{3.24}$$

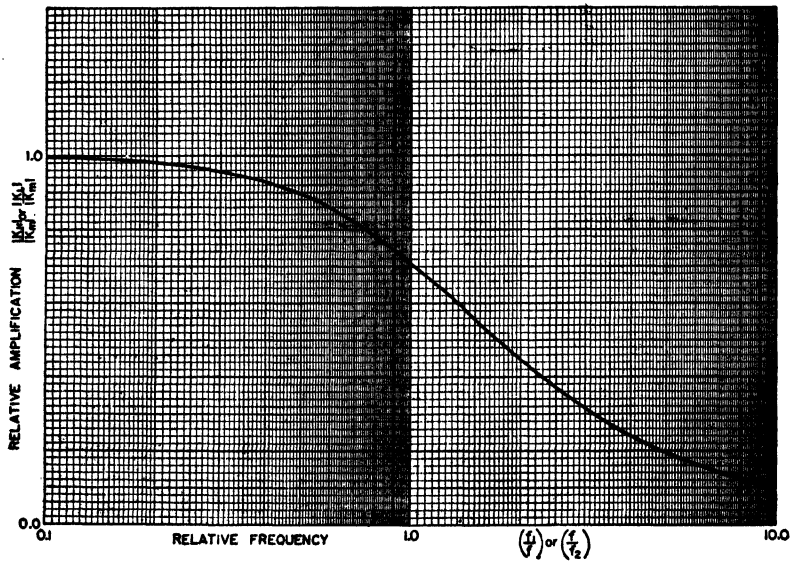


Fig. 3.8. Resistance-coupled amplifier universal amplification chart: relative amplification vs. relative frequency.

Substitution of these expressions into the gain and phase shift equations yields

$$|K_L| = \frac{|K_m|}{\sqrt{1 + (f_1/f)^2}} \tag{3.25}$$

$$\tan \theta_L = \frac{f_1}{f} \tag{3.26}$$

$$|K_H| = \frac{|K_m|}{\sqrt{1 + (f/f_2)^2}} \tag{3.27}$$

$$\tan \theta_H = -\frac{f}{f_2} \tag{3.28}$$

These last four equations are in an exceptionally convenient form since they are completely general, that is, they do not have terms peculiar to any one amplifier. If the relative voltage amplification $|K|/|K_m|$ as well as the relative phase shifts are plotted as functions of relative frequency (f_1/f) and (f/f_2), the universal amplification curves of Figs. 3.8 and 3.9 are obtained. The data required for these figures are given in Table 1.

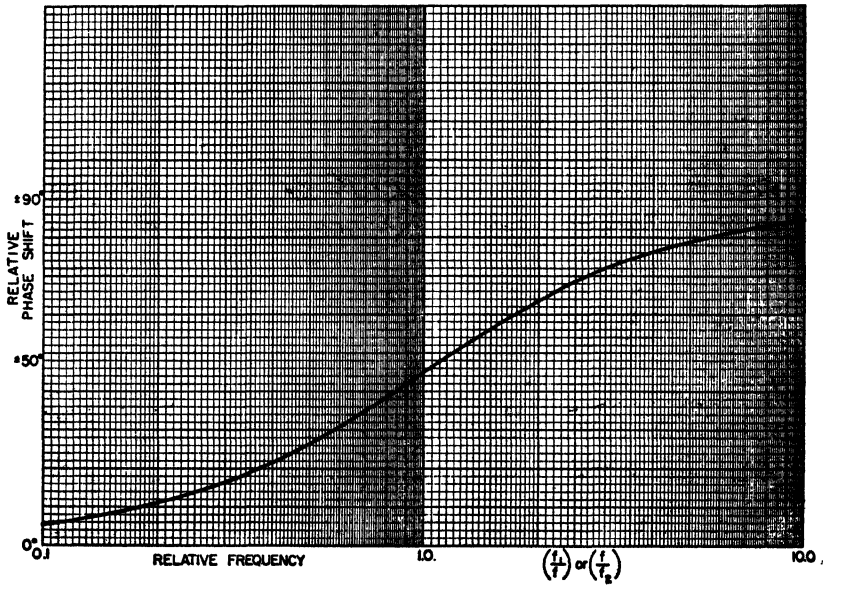


Fig. 3.9. Resistance-coupled amplifier universal amplification chart. Relative phase vs. relative frequency.

The inherent usefulness of these curves is not to be underestimated, but their use sometimes leads to misconceptions about the characteristics of amplifiers. The confusion arises since frequency *ratio*, not frequency, is being plotted. The actual frequency f is in the numerator of the ratio in the high-frequency case, but in the denominator in the low-frequency case. Thus, it is to be expected that the high-frequency gain and phase shift characteristics will differ appreciably from those at the low-frequency end when the operating frequency itself is the variable. The essential dissimilarity is illustrated by Fig. 3.10.

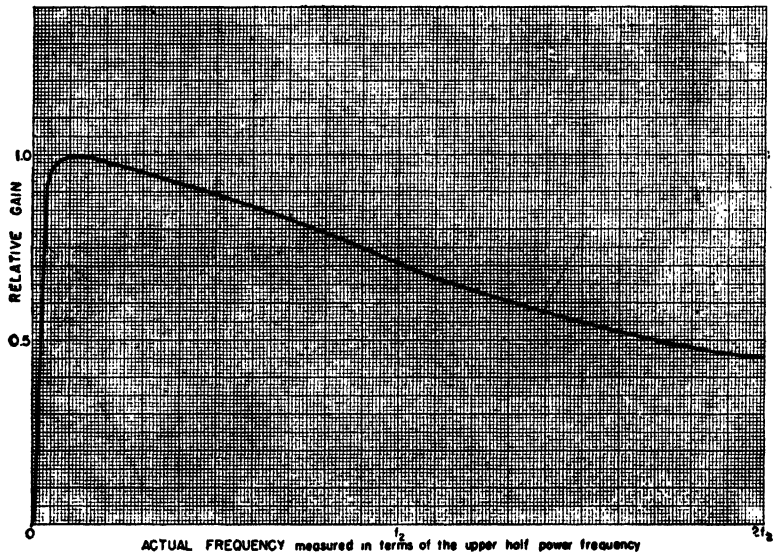


Fig. 3.10. Amplification characteristic of a resistance-coupled amplifier plotted on a linear frequency scale.

3.4 Special Case of the Pentode

The pentode was used almost exclusively in video amplifiers during the second World War since its input capacitance was considerably less than for the then current triodes. This permitted wide band amplification to be obtained with greater ease. Subsequent developments and improvements in triodes were made to such a substantial degree that they are making a comeback into the field.

TABLE 1

UNIVERSAL AMPLIFICATION CHART DATA

Low Frequencies			Relative Amplification	High Frequencies		
f_1/f	Phase Shift (+)			Phase Shift (−)		f/f_2
	deg	min		deg	min	
0.01	0	35	1.000	0	35	0.01
0.1	5	43	0.995	5	43	0.1
0.2	11	19	0.981	11	19	0.2
0.3	16	42	0.957	16	42	0.3
0.4	21	48	0.926	21	48	0.4
0.5	26	34	0.894	26	34	0.5
0.6	30	58	0.858	30	58	0.6
0.7	35	00	0.819	35	00	0.7
0.9	41	59	0.743	41	59	0.9
1.0	45	00	0.707	45	00	1.0
1.2	50	12	0.640	50	12	1.2
1.5	56	19	0.555	56	19	1.5
2.0	63	26	0.447	63	26	2.0
5.0	78	41	0.196	78	41	5.0
10.0	84	18	0.0995	84	18	10.0
100	89	26	0.0100	89	26	100

When pentodes are used, several valid approximations may be invoked which will considerably simplify the equations of Art. 3.3. In general, in a wide-band pentode amplifier,

$$r_p \approx 1 \text{ megohm} \qquad R_g \approx \tfrac{1}{2} \text{ megohm}$$
$$R_L \approx 10 \text{ to } 20 \text{ kilohms}$$

Hence, the gain equation and the expressions for the half-power frequencies reduce to

$$|K_m| = g_m R_L \tag{3.29}$$

$$\omega_2 = \frac{1}{R_L C_s} \tag{3.30}$$

$$\omega_1 = \frac{1}{R_g C_c} \tag{3.31}$$

3.5 Sources of Distortion in Resistance-Coupled Amplifiers

Examination of the transmission characteristics of the resistance-coupled amplifier, as shown in Figs. 3.8, 3.9, and 3.10, point up the fact that the conditions for no distortion, previously listed, are not met at either the high- or low-frequency ends of the amplifier pass band. It is apparent that both amplitude and delay distortion are introduced in both parts of the spectrum.

At the high-frequency end the distortion is introduced principally as a result of the presence of the shunt capacitances, both distributed, due to the capacitances to ground of the leads and circuit elements, as well as the tube interelectrode capacitances. Due to the Miller effect, the input capacitance to the second stage is altered by the gain of that stage according to the relationship

$$C_{\text{in}} = C_{gk} + C_{gp}\,(|K| + 1)$$

The amplitude distortion results from the increased shunting effect of the total shunt capacitance as the frequency is made larger. At the same time it causes the phase to shift nonlinearly. The transmission characteristics are essentially the same as an R-C circuit of the form shown in Fig. 3.11.

At the low-frequency end, the amplitude distortion results from the nonlinear change in voltage-dividing action between the coupling

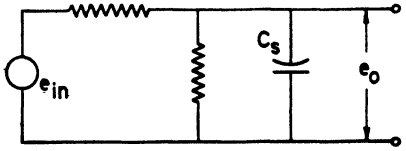


Fig. 3.11. Prototype high-frequency equivalent plate circuit.

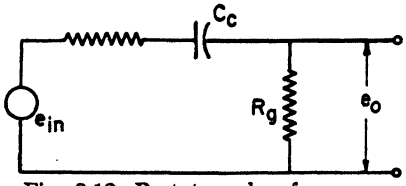


Fig. 3.12. Prototype low-frequency equivalent plate circuit.

condenser and the grid leak resistance. The transfer characteristic is essentially the same as would be obtained from the circuit of Fig. 3.12.

The essential dissimilarities between these two circuits further support the contention advanced in the preceding article that the nature and extent of the distortion introduced at the extreme ends of the pass band are different. The reason is apparent from the circuit configurations of Figs. 3.11 and 3.12. In Fig. 3.12 the output is the voltage across a resistor in a series R-C circuit. Since the circuit of Fig. 3.11 could be Thevenized to a simple series R-C circuit, it is apparent that the output voltage is the voltage across the condenser. Consequently, the transmission properties of the two circuits are bound to differ appreciably.

3.6 Methods of Extending Bandwidth without Compensation

Since the upper half-power frequency for a pentode video amplifier is given by

$$\omega_2 = \frac{1}{R_L C_s}$$

it is evident that distortion at the high-frequency end can be reduced by decreasing either the plate load resistance or the total shunt capacitance. The shunt capacitance may be minimized as follows: The coupling condenser, plate load resistor, and grid leak should be located *away* from the chassis; all signal leads should be as short as possible; by-pass condensers should be mounted as close to the tube socket as possible. A further reduction may be made by proper tube selection so that low interelectrode capacitances are obtained. In general, pentodes are used because their input capacitance is very much less than for triodes.

Having done all the minimizing of shunt capacitance possible, the other alternative remains, decreasing the plate load resistance. However, since the gain is given by

$$|K_m| = g_m R_L$$

a practical limit is reached when the reduction in R_L causes the voltage amplification to drop below that required for effective operation. When this limit is reached, the only remaining alternative is the insertion of a compensating circuit. Circuits performing this function will be discussed subsequently.

At the low-frequency end

$$\omega_1 = \frac{1}{R_g C_c}$$

and it is apparent that the bandwidth may be extended by increasing the size of either the coupling condenser or the grid leak resistance. In both cases there are definite practical limitations upon the maximum values which these parameters may have. For example, there is no point in making the grid leak resistance any larger than the grid-to-cathode leakage resistance. However, the upper limit on the value of R_g that can be used with a specific tube is governed primarily by the steady component of grid current that exists due to leakage, ionization of the residual gas, and primary and secondary emission from the grid. Generally speaking, R_g may be larger for small tubes because the grid surface is smaller and the number of ions collected is less. When R_g is too large, the voltage drop produced in it by the grid current changes the net grid bias and can move the point of operation into a nonlinear region. Consequently, the manufacturer usually specifies a maximum permissible value for the grid leak of any tube.

As the coupling capacitance is increased, the leakage conductance increases, producing d-c coupling to the grid of the succeeding tube, thus changing the bias to an improper value. Furthermore, as the capacitance is increased, the physical size of the condenser increases, thus increasing the stray capacitance to ground and reducing the upper half-power frequency.

If the maximum practical values of grid leak and coupling capacitance do not extend the lower half-power frequency to the desired point, some compensating circuit must be introduced.

3.7 High-Frequency Compensation—Shunt Peaking

When the bandwidth and gain requirements become so extreme that they cannot be met satisfactorily by the methods outlined in the preceding article, it becomes necessary to introduce circuits, or circuit elements, which will compensate for the undesirable effects, producing a net improvement in either or both of the gain and phase shift characteristics. Furthermore, the use of compensating circuits will, in general, allow a higher value of mid-frequency gain for the same over-all bandwidth, an obvious advantage.

A qualitative idea of the character of the circuit required for compensation may be obtained from recognition of the fact that the gain of the amplifier is proportional to the impedance in the plate circuit. At the high-frequency end, the impedance, and thus the gain, drops rapidly due to the shunting effect of C_s. To compensate for this effect, it would be necessary to add in another circuit element with a rising frequency characteristic which will cause the total impedance in the plate circuit to remain substantially constant over an appreciably larger frequency range.

Compensation at the high-frequency end may be accomplished by any one, or combination, of several available methods. However, the simplest in operation and concept is produced by the insertion of an inductance in series with the plate load resistance. The effect of this inductance, when it has the proper value, is to "peak" the amplifier gain vs. frequency curve at the high frequency end. Consequently, it is frequently called a "peaking inductance." Furthermore, in the a-c equivalent circuit, it is in shunt with the distributed wiring and tube interelectrode capacitances. Hence, this type of compensation is commonly referred to as "shunt peaking."

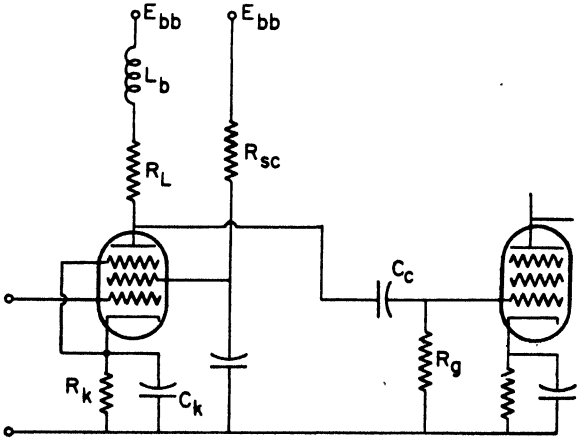


Fig. 3.13. Pentode video amplifier using shunt-peaking high-frequency compensation.

A typical shunt-peaked pentode amplifier is shown in Fig. 3.13. L_b is the compensating inductance. The equivalent plate circuit for the amplifier is given in Fig. 3.14 in which the capacitance C_s repre-

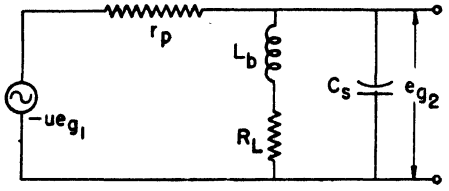


Fig. 3.14. Approximate high-frequency equivalent plate circuit for a video amplifier employing shunt peaking.

sents the sum of the input, output, and wiring shunt capacitances. That is

$$C_s = C_0 + C_i + C_w$$

As pointed out in the article on the special case of the pentode,

$$R_g \gg R_L \quad \text{and} \quad r_p \gg R_L$$

Let Z_L designate the equivalent impedance of the two parallel branches in the circuit of Fig. 3.14. Thus, according to Eq. (3.14)

$$|K| \approx \frac{g_m|Z_L|}{1 + (|Z_L|/r_p)}$$

But $Z_L \ll r_p$, and the expression for the gain becomes

$$|K| \approx g_m|Z_L| \tag{3.32}$$

But

$$Z_L = \frac{(R_L + j\omega L_b)\left(-\frac{j}{\omega C_s}\right)}{R_L + j\left(\omega L_b - \frac{1}{\omega C_s}\right)} \tag{3.33}$$

which reduces to

$$Z_L = \frac{R_L + j\omega L_b}{(1 - \omega^2 C_s L_b) + j\omega C_s R_L} \tag{3.34}$$

This may be rationalized and the following equation obtained:

$$Z_L = \frac{R_L - j(\omega C_s R_L^2 - \omega L_b + \omega^3 C_s L_b^2)}{(1 - \omega^2 C_s L_b)^2 + (\omega C_s R_L)^2} \tag{3.35}$$

Let the letter N designate the ratio between the reactance of the compensating inductance at the upper half-power frequency of the uncompensated stage and the plate load resistance. That is

$$N = \frac{\omega_2 L_b}{R_L} \tag{3.36}$$

Solving for the value of the compensating inductance yields

$$L_b = \frac{NR_L}{\omega_2} \tag{3.37}$$

It was previously shown that, for a pentode, the upper half-power frequency is given by

$$\omega_2 = \frac{1}{R_L C_s}$$

or, solving for the total shunt capacitance,

$$C_s = \frac{1}{\omega_2 R_L} \tag{3.38}$$

Substituting Eqs. (3.37) and (3.38) into Eq. (3.35) produces the following result:

$$Z_L = \frac{R_L - j\left(\frac{\omega R_L^2}{\omega_2 R_L} + \frac{\omega^3 N^3 R_L}{\omega_2^3 R_L} - \frac{\omega N R_L}{\omega_2}\right)}{\left(1 - \frac{\omega^2 N R_L}{\omega_2^2 R_L}\right)^2 + \left(\frac{\omega^2 R_L^2}{\omega_2^2 R_L^2}\right)} \tag{3.39}$$

Cancelling out terms leaves

$$Z_L = \frac{\left\{1 - j\left[\left(1 - N\right)\frac{f}{f_2} + N^2\left(\frac{f}{f_2}\right)^3\right]\right\} R_L}{\left[1 - N\left(\frac{f}{f_2}\right)^2\right]^2 + \left(\frac{f}{f_2}\right)^2} \tag{3.40}$$

When substituted in the gain equation this yields

$$K = -g_m R_L \frac{1 - j\left[\left(1 - N\right)\frac{f}{f^2} + N^2\left(\frac{f}{f^2}\right)^3\right]}{\left[1 - N\left(\frac{f}{f^2}\right)^2\right]^2 + \left[\frac{f}{f^2}\right]^2} \tag{3.41}$$

The magnitude of the voltage amplification is then

$$|K| = g_m R_L \frac{\sqrt{1 + \left[(1 - N)\frac{f}{f^2} + N^2\left(\frac{f}{f_2}\right)^3\right]^2}}{\left[1 - N\left(\frac{f}{f_2}\right)^2\right]^2 + \left[\frac{f}{f_2}\right]^2} \tag{3.42}$$

and the relative phase shift is given by

$$\tan \Theta = -\left[(1 - N)\frac{f}{f^2} + N^2\left(\frac{f}{f_2}\right)^3\right] \tag{3.43}$$

The value of N for which the gain at the upper half-power frequency of the uncompensated stage is equal to its mid-frequency value $g_m R_L$ may be obtained by setting

$$f/f_2 = 1 \qquad \text{and} \qquad |K| = g_m R_L = |K_m|$$

Thus, it is found that, for this condition,

$$N = 0.50 \tag{3.44}$$

The compensation obtained when $N = 0.50$ is generally referred to as the compensation for *best frequency response*, since the voltage amplification does not deviate appreciably from the mid-frequency value for frequencies up to and including the upper half-power frequency of the uncompensated stage. This is illustrated by Table 2 and Fig. 3.15. However, it is evident that this calculation does not insure *uniform* voltage amplification at frequencies below the upper half-power frequency for the uncompensated stage.

It will be noted from Fig. 3.15 and from the data in Table 2 that

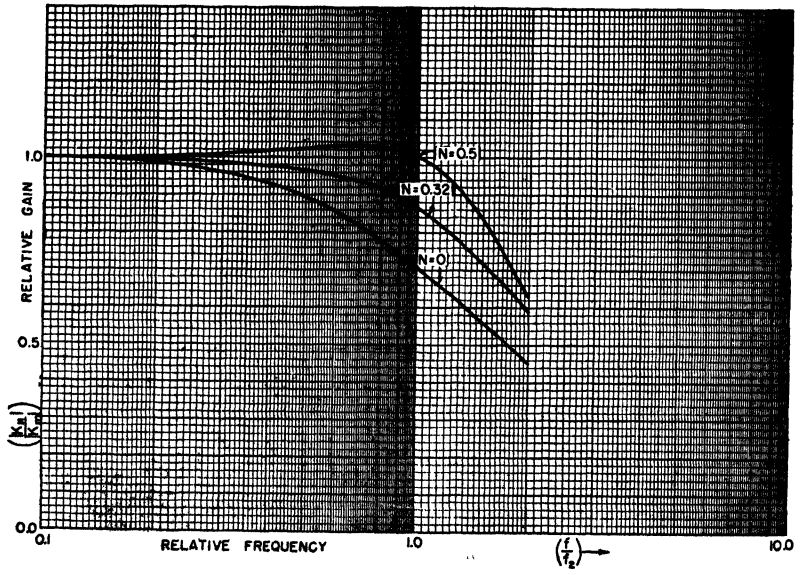


Fig. 3.15. Shunt peaked resistance-coupled amplifier: relative phase vs. relative frequency.

the value of $N = 0.50$ does not result in the phase shift being proportional to frequency within the amplifier pass band. It does, however, represent an improvement over the phase shift characteristic for the uncompensated stage.

The value of N for minimum phase distortion has been found to be 0.32. This results in the data of Table 3 and the curves of Figs. 3.15 and 3.16. By placing a straightedge along the phase characteristic, it will be observed that it is very linear over the bulk of the characteristic.

However, when $N = 0.32$, the gain at the half-power frequency of the uncompensated stage is only 87.7 per cent of its mid-frequency value. Since this decrease in gain is too large for most applications, a compromise value of $N = 0.44$ should be used when it is desired to improve the amplitude and phase characteristics simultaneously. When $N = 0.44$ the voltage amplification at the uncompensated upper half-power frequency is 95.3 per cent of its maximum value. This variation is well within a 1 db limit.

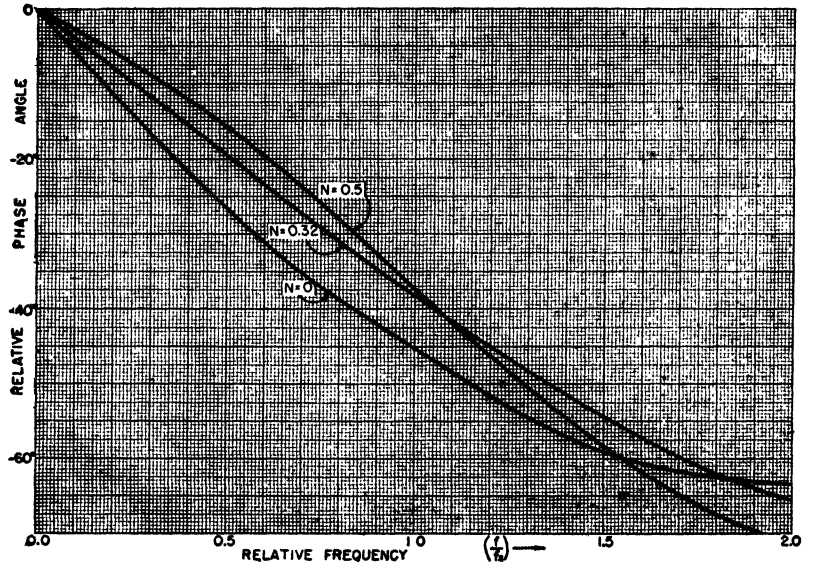


Fig. 3.16. Shunt peaked resistance-coupled amplifier: relative phase vs. relative frequency.

TABLE 2

DATA FOR UNIVERSAL AMPLIFICATION CURVES

Shunt Peaking N = 0.50

Relative Frequency (f/f_2)	Relative Amplification	Relative phase shift	
		deg	min
0.01	1.000013	0	18
0.1	1.001	2	53
0.2	1.005	5	49
0.3	1.0112	8	55
0.4	1.0169	12	11
0.5	1.0220	15	41
0.6	1.0275	19	30
0.7	1.0281	23	33
0.8	1.025	27	37
0.9	1.016	32	17
1.0	1.000	36	53
1.10	0.976	41	21
1.20	0.947	45	54
1.50	0.829	57	54

TABLE 3

DATA FOR UNIVERSAL AMPLIFICATION CURVES

Shunt Peaking N = 0.32

0.3	0.987	11.7
0.4	0.977	15.6
0.5	0.966	19.4
0.7	0.935	27.1
0.8	0.913	30.8
1.0	0.877	38.0
1.5	0.720	53.8
2.0	0.587	65.3

3.8 Figure of Merit

It has been shown that the mid-frequency voltage amplification of a shunt-peaked, wide-band, pentode amplifier is given by

$$|K_m| = g_m R_L$$

But since

$$\omega_2 = \frac{1}{R_L C_s}$$

then $$R_L = \frac{1}{\omega_2 C_s}$$

This reduces the equation for the voltage amplification to

$$|K_m| = \frac{g_m}{\omega_2 C_s} \tag{3.45}$$

Now, there are generally two conflicting requirements on a wide-band amplifier: both the upper half-power frequency and the gain should be large. Consequently, the product of these two quantities can be used as a performance factor in rating amplifiers. Designating this factor as F_a, then

$$F_a = \omega_2 |K_m| = \frac{g_m}{C_s} = \frac{g_m}{C_0 + C_i + C_w} \tag{3.46}$$

F_a is generally called the amplifier "figure of merit." Since the distributed wiring capacitance C_w is the only quantity that is not determined by the tubes used in the amplifier, it is generally convenient to compare tubes in terms of a figure of merit F_t where

$$F_t = \frac{g_m}{C_0 + C_i} \tag{3.47}$$

Table 4 lists the figures of merit of several typical tubes. Because of its high figure of merit, the 6AC7 is extensively used for video voltage amplifiers, whereas the 6AG7 is widely used as a power amplifier.

TABLE 4
FIGURES OF MERIT FOR TYPICAL TUBES

$$F_t = \frac{g_m}{C_o + C_i}$$

Tube	g_m in μmhos	C_o in $\mu\mu$f	C_i in $\mu\mu$f	$F_t \times 10^6$
6AB7	5000	5	8	385
6AC7	9000	5	11	562
6AG7	7700	7.5	12.5	385
6C5	2000	25	11	55
6L6	5200	12	10	236
6SJ7	1650	7	6	127
6SK7	2000	7	6	154
6V6	4100	11	10	195
6Y6G	7100	8	15	309
6SG7	4700	7	8.5	303
7G7	4500	7	9	281

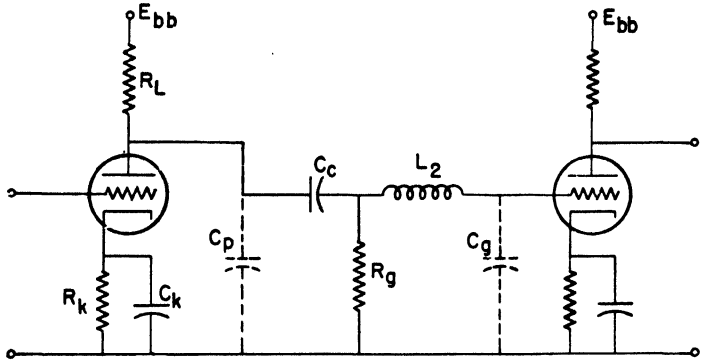


Fig. 3.17. Video amplifier using series peaking.

3.9 High-Frequency Compensation—Series Peaking

Another form of high-frequency compensation is produced when an inductance is inserted in series with the coupling condenser as shown in Fig. 3.17. This is called *series peaking*. The equivalent plate circuit is shown in Fig. 3.18. Compensation is produced due to the voltage multiplication that takes place as the resonant frequency of the tuned circuit produced by the inclusion of this inductance is approached. Another factor contributing to better performance arises since the total shunt capacitance has been separated, by the insertion of L_2, into two components, C_p and C_g, so that the load resistance is directly shunted only by C_p. Consequently, since

$$\omega_2 = \frac{1}{R_L C \text{ (in shunt with } R_L)}$$

it is possible to produce higher gains for a given bandwidth because

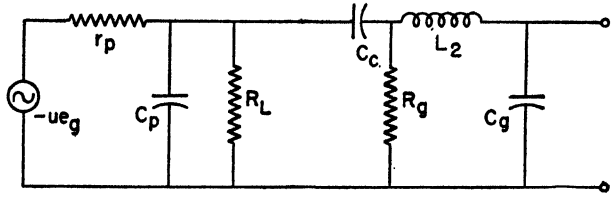


Fig. 3.18. Equivalent plate circuit of a series-peaked video amplifier.

the load resistance can be larger than is possible in shunt peaking where both C_p and C_g shunt the plate load.

Since the coupling condenser is effectively a short circuit at the frequencies under consideration, it may be omitted from the high-frequency equivalent plate circuit. Furthermore, by redrawing the circuit, it appears as shown in Fig. 3.19. The circuit between the

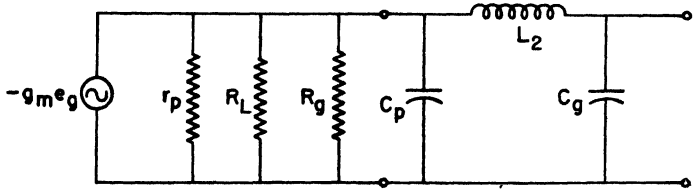


Fig. 3.19. High-frequency equivalent plate circuit of a series-peaked video amplifier.

four terminals indicated appears to be a π-section, constant-k (low-pass) filter where R_L is one termination and R_g is the other. In general, for such a filter, optimum performance is obtained when $C_g = 2C_p$. This is a rather fortunate circumstance because in most video tubes, such as the 6AC7 (see Table 4), this condition is approached rather closely since the input capacitance is generally larger than the output capacitance. Furthermore, by placing the coupling condenser on the plate side of the compensating inductance, C_p may be increased by the additional wiring capacitance. It is not desirable to simply add another physical capacitor in order to attain the two-to-one ratio. For one thing, the ratio does not affect operation to a large extent (that is, it is not critical), and secondly, the addition of a second physical condenser will only make performance worse.

Collecting the three parallel resistances into one single equivalent yields the circuit of Fig. 3.20. Application of Kirchhoff's node law at the points indicated yields for node A

$$g_m e_g = E_1\left[\frac{1}{R} + j\left(\omega C_p - \frac{1}{\omega L_2}\right)\right] + E_2\left(j\,\frac{1}{\omega L_2}\right) \tag{3.48}$$

for node B

$$0 = E_1\left(j\,\frac{1}{\omega L_2}\right) + E_2\left[\frac{1}{R} + j\left(\omega C_g - \frac{1}{\omega L_2}\right)\right] \tag{3.49}$$

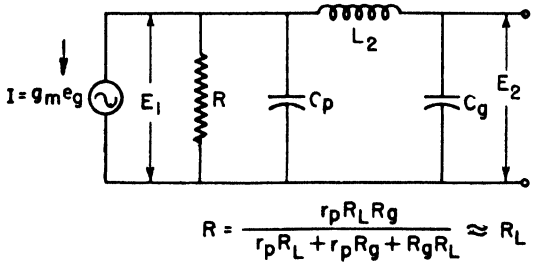


Fig. 3.20. High-frequency equivalent plate circuit of a series-peaked video amplifier.

Eliminating E_1 by solving these equations simultaneously yields

$$g_m e_g = E_2(j\omega L_2)\left[j\left(\omega C_g - \frac{1}{\omega L_2}\right)\right]\left[\frac{1}{R} + j\left(\omega C_p - \frac{1}{\omega L_2}\right)\right] + j\frac{E_2}{\omega L_2} \tag{3.50}$$

Solving for the voltage amplification

$$K = \frac{E_2}{e_g} = \frac{g_m}{\left(1 - \omega^2 L_2 C_g\right)\left[\frac{1}{R} + j\left(\omega C_p - \frac{1}{\omega L_2}\right)\right] + j\left(\frac{1}{\omega L_2}\right)} \tag{3.51}$$

Obtaining the common denominator results in

$$K = \frac{(j\omega L_2)g_m R}{(1-\omega^2 L_2 C_g)(R - \omega^2 R C_p L_2 + j\omega L_2) - R} \tag{3.52}$$

The calculation of this equation is extremely difficult in comparison with that obtained for the case of shunt peaking. Further-

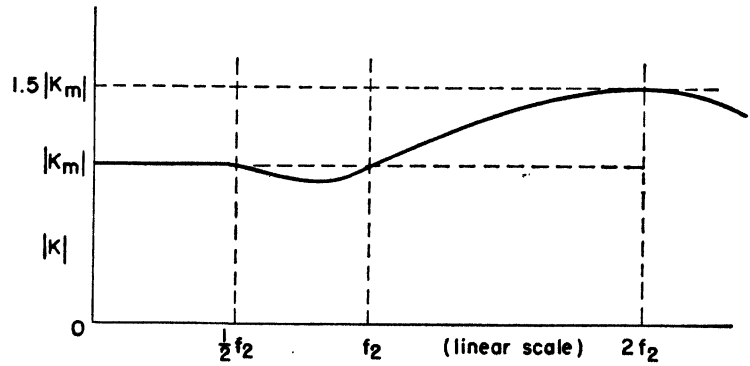


Fig. 3.21. Gain vs. frequency for series-peaked video amplifier.

more, it is not possible to tabulate results in the generalized and simple form obtained for shunt peaking.

When $C_g = 2C_p$, the gain characteristic appears as shown in Fig. 3.21. The bandwidth produced is greater than that obtained with shunt peaking, and an improvement in the phase characteristic is noted. Due to the sharper cutoff at the high-frequency end, the transient response will not be as good as for the shunt-peaked case and could produce transient oscillations which would be very detrimental to performance.

The proper design relationships have been found to be

$$R_L = \frac{1.5}{2\pi f_2 C_s} \tag{3.53}$$

$$L_2 = 0.67\, C_s R_L^2 \tag{3.54}$$

3.10 Resumé of Other Methods of High-Frequency Compensation

An alternative method of compensation which immediately suggests itself is a combination of shunt and series peaking. With proper design it can produce a nearly constant amplification characteristic up to four times the upper half-power frequency of the uncompensated stage. The proper design values have been found to be

$$R_L = \frac{1.8}{2\pi f_2 C_s} \tag{3.55}$$

$$L(\text{shunt}) = 0.12\, C_s R_L^2 \tag{3.56}$$

$$L(\text{series}) = 0.52\, C_s R_L^2 \tag{3.57}$$

Another possibility arises from consideration of the shunt peaking characteristics. It was noted that the compensating inductance that produced best compensation below the upper half-power frequency for the uncompensated stage was quite small, whereas the inductance required for best compensation beyond this half-power frequency was quite large. Thus, an inductance which increases with frequency would give better over-all compensation. Since a high Q, parallel-tuned circuit, operating slightly below resonance, has such a characteristic, it could be used in place of the peaking inductance of the shunt-compensated circuit.

Alert experimenters observed that the transmission characteristics desired for the interstage coupling networks of amplifiers were substantially the same as those produced by low-pass filters. One such

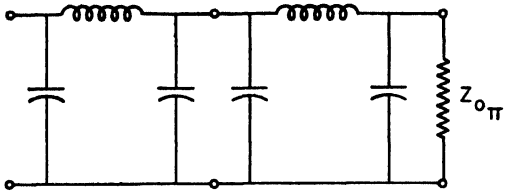


Fig. 3.22. Two-section constant-k filter.

possibility is series peaking, which has already been discussed. An extension of this idea has been developed as follows: Consider a two-section π-type low-pass constant-k filter terminated in its characteristic impedance. The circuit for this filter is shown in Fig. 3.22 and the variation of input impedance, which equals the characteristic impedance, with frequency is given in Fig. 3.23. A shunt m-derived section, when $m = 0.6$, has a characteristic impedance which is practically constant, and equal to $R = \sqrt{L/C}$, over the entire pass band as shown in Fig. 3.25. Furthermore, the input impedance of half a shunt m-derived filter, terminated in its characteristic impedance, is equal to the characteristic impedance of a constant-k π-section. Thus, if the original π-section filters are terminated in shunt m-derived half T sections, with $m = 0.6$, the circuit appears as in Fig. 3.24 and the filters are matched throughout. In practical use, the capacitive shunt arms of the constant-k sections are the circuit and interelectrode capacitances.

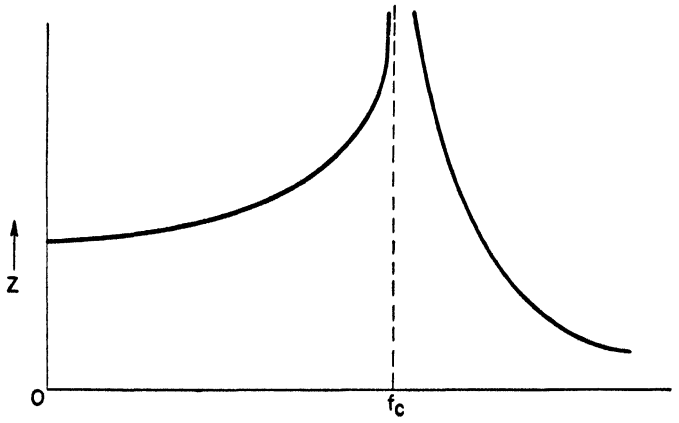


Fig. 3.23. Constant-k frequency characteristic.

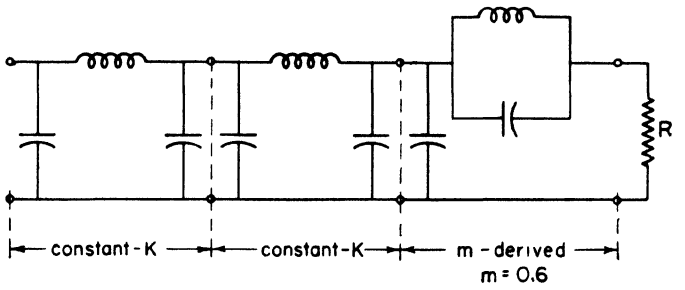


Fig. 3.24. Cascaded filter sections.

The addition of the shunt m-derived half T sections does not effectuate any change in the variation of input impedance of the composite filter as the frequency changes. However, it does allow termination with an ordinary resistance without producing a mismatch between terminals of the filter sections. Now, since the amplifier gain is given by

$$K = g_m Z_L$$

then, for effective operation, the load impedance Z_L, which is also the input impedance of the composite filter, should be constant over the entire desired pass band of the amplifier. This could be obtained by adding an additional capacitance C_n directly in shunt with the first capacitance of the first constant-k section. The input impedance of the filter is

$$Z_{in} \approx \frac{R}{\sqrt{1 - (f/f_c)^2}} \qquad \text{where} \qquad R = \sqrt{\frac{L}{C}}$$

When the condenser C_n is added, then

$$Z_L = \frac{1}{(1/Z_{in}) + j\omega C_n} = \frac{R}{\sqrt{1 - (f/f_c)^2 + j\omega C_n R}}$$

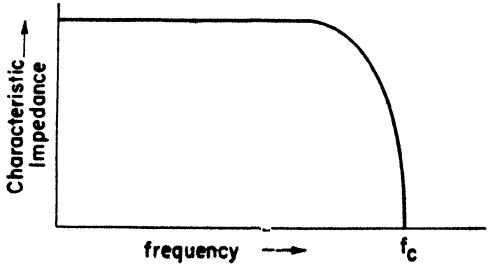


Fig. 3.25. m-derived frequency characteristic of Z_0, with (m = 0.6).

or $$|Z_L| = \frac{R}{\sqrt{1 - (f/f_c)^2 + (\omega C_n R)^2}}$$

The magnitude of the input impedance ($|Z_L|$) is independent of frequency when

$$\frac{\omega}{\omega_c} = \omega C_n R \qquad \text{or} \qquad C_n = \frac{1}{\omega_c R} = \frac{1}{2\pi f_c R}$$

Let C_π = total shunt capacitance of the π-section filter
Then let $C_n = nC_\pi$

$n = 1$ gives best amplitude response.
$n = 1.2$ gives a very nearly linear phase shift with frequency.
$n = 1.155$ is the best compromise.

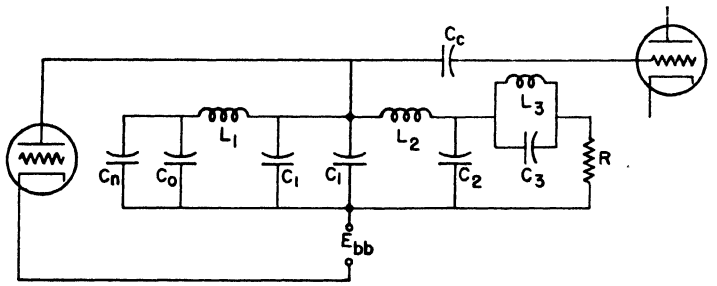


Fig. 3.26. Filter-coupled video amplifier.

The practical circuit connections are shown in Fig. 3.26. The appropriate design relationships have been found to be

$$\left.\begin{aligned}
C_0 &= C_1 = \frac{1}{\pi f_c R} \\
C_2 &= \frac{1}{2}\left(\frac{1}{\pi f_c R}\right) + \frac{m}{2}\left(\frac{1}{\pi f_c R}\right) = \frac{C_0}{2}(1 + m) \\
C_3 &= \left(\frac{1 - m^2}{4m}\right)\left(\frac{2}{\pi f_c R}\right) = C_0\left(\frac{1 - m^2}{2m}\right) \\
C_n &= n\left(\frac{3C_0}{2}\right) \\
L_1 &= L_2 = \frac{R}{\pi f_c} \\
L_3 &= \frac{mL_1}{2} \\
m &= 0.6
\end{aligned}\right\} \qquad (3.58)$$

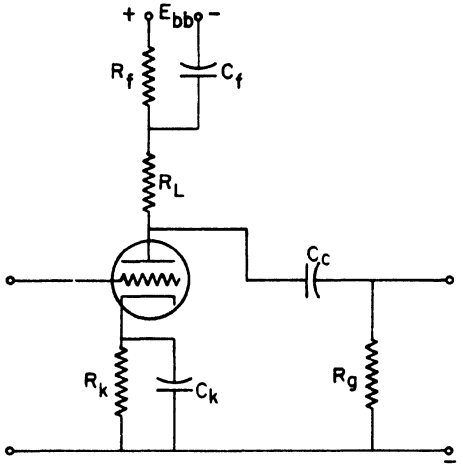


Fig. 3.27(a). Low-frequency compensation of a video amplifier.

3.11 Low-Frequency Compensation

The drooping of the voltage amplification characteristic and the nonlinearity of the phase shift at the low-frequency end of the amplifier pass band can be offset by inserting a parallel combination of resistance and capacitance in series with the plate load resistance of the tube. Due to this circuit connection, the impedance in series with the plate of the tube increases with decreasing frequency, thus tending to offset the drop in amplifier gain caused by the grid coupling circuit. Typical circuit connections are shown in Fig. 3.27(a) and the equivalent plate circuit is given in Fig. 3.27(b).

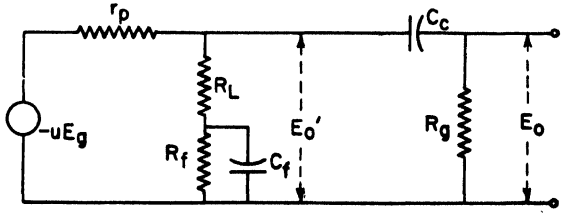


Fig. 3.27(b). Equivalent plate circuit of low-frequency compensated video amplifier.

Let Z'_L designate the total impedance in series with the plate, that is,

$$Z'_L = R_L - j\frac{R_f X_{Cf}}{R_f - jX_{Cf}} \tag{3.59}$$

Neglecting the effect of the grid coupling circuit, the intermediate voltage amplification may be written as

$$K'_L = \frac{E'_0}{E_g} = -\frac{g_m Z'_L}{1 + (Z'_L/r_P)} \tag{3.60}$$

Assuming that a pentode is used, and this is frequently the case for video amplifiers, then $r_p \gg Z'_L$ and the expression for the intermediate amplification reduces to

$$K'_L = -g_m Z'_L \tag{3.61}$$

Now, this ideal amplification is cut down by the grid coupling network $C_c - R_g$ so that the actual amplification is

$$K_L = \frac{E_0}{E_g} = K'_L\left(\frac{R_g}{R_g - jX_{Cc}}\right) \tag{3.62}$$

Substituting for K'_L from Eq. (3.61) yields

$$K_L = -g_m R_g\left(\frac{Z'_L}{R_g - jX_{Cc}}\right) \tag{3.63}$$

Substitute Eq. 3.59 for the impedance Z'_L, obtaining

$$K_L = -g_m R_g\left(\frac{R_L - j\dfrac{R_f X_{Cf}}{R_f - jX_{Cf}}}{R_g - jX_{Cc}}\right) \tag{3.64}$$

Reduction to a common denominator yields

$$K_L = -g_m R_g \frac{1 - jX_{Cf}(R_f + R_L)}{(R_f - jX_{Cf})(R_g - jX_{Cc})} \tag{3.65}$$

Divide numerator and denominator through by the product $R_g R_L R_f$.

$$K_L = -g_m R_L \frac{1 - jX_{Cf}\left(\dfrac{R_f + R_L}{R_f R_L}\right)}{\left(1 - j\dfrac{X_{Cf}}{R_f}\right)\left(1 - j\dfrac{X_{Cc}}{R_g}\right)} \tag{3.66}$$

Then replace the bracketed quantity in the numerator by R', that is

$$R' = \frac{R_f R_L}{R_f + R_L} \tag{3.67}$$

so that the voltage amplification may be written as

$$K_L = -g_m R_L \frac{\left(1 - j\frac{X_{Cf}}{R'}\right)}{\left(1 - j\frac{X_{Cf}}{R_f}\right)\left(1 - j\frac{X_{Cc}}{R_g}\right)} \tag{3.68}$$

This may be rewritten directly as

$$K_L = -g_m R_L \frac{\left(1 - \frac{j}{\omega R' C_f}\right)}{\left(1 - \frac{j}{\omega R_f C_f}\right)\left(1 - \frac{j}{\omega R_g C_c}\right)} \tag{3.69}$$

From experience it has been found that optimum compensation occurs when the following time constants are equal.

$$R'C_f = R_g C_c \tag{3.70}$$

Using this equality, the expression for the voltage amplification of the compensated amplifier becomes

$$K_L = -\frac{g_m R_L \,|\underline{\Theta_L}}{\sqrt{1 + (1/\omega R_f C_f)^2}} = \frac{K_m}{(1 - j/\omega R_f C_f)} \tag{3.71}$$

In polar form, this may be written as

$$|K_L| = \frac{|K_m|}{\sqrt{1 + (1/\omega R_f C_f)^2}} \tag{3.72}$$

$$\Theta_L = 180^0 + \tan^{-1}\left(\frac{1}{\omega R_f C_f}\right) \tag{3.73}$$

These expressions are fairly accurate as long as fixed bias is used. When cathode bias is employed, there will be some cathode circuit feedback which tends to reduce the low-frequency gain and shift the phase in the same direction as that caused by the grid coupling circuit. Additional compensation is necessitated. This is discussed in detail in Art. 3.13.

3.12 Effects of Multistaging

When a series of identical amplifier stages are connected in cascade, the bandpass of each individual stage must be increased if the over-all bandwidth of the amplifier is to remain constant. This becomes an important consideration since multistaging is generally required in television video amplifiers.

In order to keep the nomenclature straight, let

f_2 = upper half-power frequency for a single stage
f_2' = overall upper half-power frequency for n identical stages connected in cascade
$|K_s|$ = absolute gain of a single stage
$|K_n|$ = absolute gain of n identical stages connected in tandem

It was previously shown that, for a single stage, the high-frequency gain equation for the uncompensated amplifier may be written as

$$|K_s| = \frac{|K_s|_m}{[1 + (f/f_2)^2]^{1/2}} \tag{3.74}$$

The total gain of the complete amplifier is simply the product of all of the individual stage gains. Since the stages were assumed to be identical, then

$$|K_n| = |K_s|^n \tag{3.75}$$

Now, substituting Eq. (3.74) into (3.75) produces

$$|K_n| = \frac{(|K_s|_m)^n}{[1 + (f/f_2)^2]^{n/2}} \tag{3.76}$$

The relationship between the mid-frequency gains is, of course,

$$(|K_s|_m)^n = |K_n|_m$$

So,

$$|K_n| = \frac{|K_n|_m}{[1 + (f/f_2)^2]^{n/2}} \tag{3.77}$$

At the over-all upper half-power frequency f'_2 for the complete cascaded amplifier of n stages,

$$|K_n| = \frac{|K_n|_m}{\sqrt{2}} \tag{3.78}$$

When substituted in Eq. (3.77) it produces

$$\frac{|K_n|_m}{\sqrt{2}} = \frac{|K_n|_m}{[1 + (f_2'/f_2)^2]^{n/2}} \tag{3.79}$$

Since the numerators are equal, then the denominators must likewise be equal, or

$$\left[1 + \left(\frac{f_2'}{f_2}\right)^2\right]^{n/2} = \sqrt{2} \tag{3.80}$$

Square both sides of Eq. (3.80) and then take the nth root.

$$1 + \left(\frac{f_2'}{f_2}\right)^2 = 2^{1/n} \tag{3.81}$$

Solve for f'_2, obtaining

$$f_2' = f_2\sqrt{2^{1/n} - 1} \tag{3.82}$$

Following exactly the same procedure at the low-frequency end yields

$$f_1' = \frac{f_1}{\sqrt{2^{1/n} - 1}} \tag{3.83}$$

The radical can be evaluated for different numbers of stages as given in Table 5.

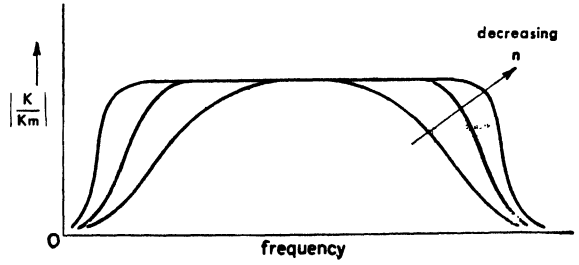


Fig. 3.28. Effect of multistaging on overall bandwidth.

The effects on the bandwidth are shown in Fig. 3.28. Consequently, from cost considerations and construction difficulties caused by the higher bandwidths per stage required, it is desirable to keep the number of stages to a minimum. Further, the results indicate

TABLE 5

VALUES OF THE FACTOR $\sqrt{2^{1/n} - 1}$

n	$\sqrt{2^{1/n} - 1}$	n	$\sqrt{2^{1/n} - 1}$
1	1.000	6	0.350
2	0.643	7	0.323
3	0.510	8	0.301
4	0.435	9	0.283
5	0.387	10	0.268

the desirability of employing compensation wherever possible, since stage-by-stage compensation will allow the designer to have each stage operating with the upper and lower half-power frequencies intended for the cascaded amplifier. Then by compensating each stage separately (using $N = 0.50$ with shunt peaking, for example), the total bandpass can be brought up to that desired without appreciable loss in gain. However, care must be exercised to insure that the sharper cutoff caused by multistaging the compensated amplifiers does not produce oscillations during the transmission of pulses.

Multistaging has no appreciable effect upon the *shape* of the time delay characteristic. The effect is simply to increase the total delay since if

φ_s = phase shift in a single stage
φ_n = total phase shift in n stages

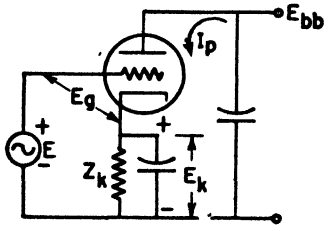
then $\varphi_n = n\varphi_s$

3.13a Effect of Cathode Degeneration—No Compensation

All of the material presented up to this point has been based upon the assumption that the cathode resistor was perfectly bypassed by the condenser C_k. This is not ordinarily true in video amplifiers for two reasons:

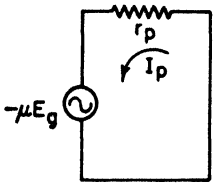
(1) The very low frequencies which must be considered.
(2) The fact that high-transconductance tubes are customarily used, thereby tending to increase the degenerative effect of the cathode circuits.

The low-frequency compensating network, if one is used, must



(a)

Fig. 3.29(a). Circuit for determining the effect of cathode degeneration on gm.



(b)

Fig. 3.29(b). Equivalent plate circuit of Fig. 3.29(a).

compensate for the distortion introduced by the cathode circuit as well as the grid coupling circuit if it is to perform effectively.

The condition that exists when the reactance of the cathode bypass condenser is not negligible is most conveniently treated as a simple case of degenerative feedback. Figure 3.29(a) will be used to show how the tube transconductance is affected by this degeneration. The equivalent plate circuit is shown in Fig. 3.29(b). The grid voltage, referred to the cathode, is

$$-E_g = E - E_k$$

Summing up voltage drops in the equivalent circuit yields

$$-\mu E_g = u(E - E_k) = I_p r_p$$

and, solving for I_p,

$$I_p = \frac{\mu}{r_p}(E - E_k) = g_m(E - E_k)$$

But, $E_k = I_p Z_k$

So $I_p = g_m E - (g_m Z_k) I_p$

Or $I_p(1 + g_m Z_k) = g_m E$

The effective transconductance, g_m', is then

$$g_m' = \frac{g_m}{1 + g_m Z_k} = \frac{I_p}{E}$$

Thus, it is apparent that the effective transconductance of the tube is reduced from its normal value by the ratio of $1/(1 + g_m Z_k)$. For many tubes, though there are a large number of exceptions, $g_m Z_k \approx 1$. Consequently, the effective transconductance may be of the order of one half its normal value.

A similar analysis of the plate circuit gives an effective plate resistance r_p' of

$$r_p' = r_p(1 + g_m Z_k)$$

showing that it is increased in the same proportion that the transconductance is decreased.

In audio amplifiers, the plate load and grid leak are generally so large compared to the normal plate resistance that the change due to cathode degeneration does not produce an appreciable reduction in amplification. However, in video amplifiers, the plate load is ordinarily smaller than the variational plate resistance and

this is an important effect. Gain reductions of the order of 10 db can result.

In Art. 3.4, for the special case of the pentode, it was shown that the low-frequency gain equation could be written as

$$K_L = \frac{g_m R_L}{1 - j(f_1/f)} \qquad \text{where} \qquad f_1 = \frac{1}{2\pi R_g C_c} \qquad g_m R_L = |K_m|$$

when perfect cathode by-passing was assumed. Now, considering the possibility of cathode degeneration, the modified value of the transconductance, g_m', must be used, so that the new low-frequency equation is

$$K_L' = \frac{g_m' R_L}{1 - j\left(\frac{f_1}{f}\right)} = \frac{g_m R_L}{(1 + g_m Z_k)\left(1 - j\frac{f_1}{f}\right)} = \frac{K_m}{(1 + g_m Z_k)\left(1 - j\frac{f_1}{f}\right)}$$

Since the cathode circuit consists of resistance and capacitance in parallel, then the cathode impedance is

$$Z_k = \frac{1}{\left(\frac{1}{R_k}\right) + j\omega C_k} = \frac{R_k}{1 + j\omega C_k R_k} = \frac{R_k}{1 + j\left(\frac{f}{f_k}\right)}$$

where

$$f_k = \frac{1}{2\pi R_k C_k}$$

Substituting into the gain equation yields

$$K_L' = \frac{K_m}{\left[1 + \frac{g_m R_k}{1 + j\left(\frac{f}{f_k}\right)}\right]\left[1 - \frac{f_1}{f}\right]}$$

$$= \left[\frac{1 + j\left(\frac{f}{f_k}\right)}{(1 + g_m R_k) + j\left(\frac{f}{f_k}\right)}\right]\left[\frac{K_m}{1 - j\left(\frac{f_1}{f}\right)}\right]$$

Let $f_3 = f_k(1 + g_m R_k)$ or $(1 + g_m R_k) = f_3/f_k$

Then, multiply numerator and denominator of the gain equation through by $-j(f_k/f)$ to obtain

$$K_L' = K_m \frac{\left(1 - j\frac{f_k}{f}\right)}{\left(1 - j\frac{f_3}{f}\right)\left(1 - j\frac{f_1}{f}\right)}$$

Or, in terms of the magnitude and phase angle

$$|K_L'| = \frac{K_m\sqrt{1 + \left(\frac{f_k}{f}\right)^2}}{\sqrt{1 + \left(\frac{f_3}{f}\right)^2}\sqrt{1 + \left(\frac{f_1}{f}\right)^2}}$$

$$\theta_L' = \tan^{-1}\left(\frac{f_1}{f}\right) + \tan^{-1}\left(\frac{f_3}{f}\right) - \tan^{-1}\left(\frac{f_k}{f}\right)$$

Note that the cathode circuit, through the (f_3/f) term, tends to reduce the gain and increase the phase advancement just as the coupling circuit does through the term (f_1/f). The effect of (f_k/f) is more than offset by (f_3/f) because f_3 is always larger than f_k.

In audio-amplifier design, a rule of thumb frequently used in calculating the proper size for the cathode by-pass condenser C_k is to let the reactance of C_k, at the frequency f_1, be $\frac{1}{10}R_k$. That is

$$\frac{1}{\omega_1 C_k} = \frac{R_k}{10} \quad \text{or} \quad f_1 = 10f_k$$

The additional phase shift introduced by the cathode circuit, using such values, is perfectly allowable for audio amplifiers, but is excessive for video amplifiers in applications such as television. The increase in phase shift can be kept small only by keeping f_k, and hence f_3, small. This requires very large values for C_k.

3.13b Effect of Cathode Degeneration—Low-Frequency Compensation

Inasmuch as the low-frequency compensating network discussed in Art. 3.11 improves the phase as well as the amplification characteristic, the information given there must be revised to include the effect of cathode degeneration if proper compensation is to be obtained.

In Eq. (3.69), the expression for the impedance in the plate circuit of the low-frequency compensated amplifier was found to be

$$Z_L = \frac{\omega R_L\left(\omega - j\frac{1}{R'C_f}\right)}{\left(\omega - j\frac{1}{R_f C_f}\right)\left(\omega - j\frac{1}{R_g C_c}\right)} \quad \text{where} \quad R' = \frac{R_L R_f}{R_L + R_f}$$

As a matter of definition, let

$$\omega_1 = \frac{1}{R_g C_c} \quad \text{or} \quad f_1 = \frac{1}{2\pi R_g C_c}$$
$$\omega_4 = \frac{1}{R' C_f} \quad \text{or} \quad f_4 = \frac{1}{2\pi R' C_f}$$
$$\omega_5 = \frac{1}{R_f C_f} \quad \text{or} \quad f_5 = \frac{1}{2\pi R_f C_f}$$

Substituting these relationships and dividing numerator and denominator by $(2\pi f)^2$ yields

$$Z_L = \frac{R_L\left(1 - j\frac{f_4}{f}\right)}{\left(1 - j\frac{f_5}{f}\right)\left(1 - j\frac{f_1}{f}\right)}$$

The voltage amplification, when cathode degeneration is considered, is

$$K_L' = g_m' Z_L = \frac{g_m R_L\left(1 - j\frac{f_4}{f}\right)}{(1 + g_m Z_k)\left(1 - j\frac{f_5}{f}\right)\left(1 - j\frac{f_1}{f}\right)}$$

In the preceding article it was found, since

$$Z_k = \frac{R_k}{1 + j\omega R_k C_k} \quad \text{that} \quad \frac{1}{1 + g_m R_k} = \frac{\left(1 - j\frac{f_k}{f}\right)}{\left(1 - j\frac{f_3}{f}\right)}$$

Substituting for Z_k, and letting $K_m = g_m R_L$, yields

$$K_L' = \frac{K_m}{\left(1 - j\frac{f_1}{f}\right)\left(1 - j\frac{f_5}{f}\right)} \frac{\left(1 - j\frac{f_4}{f}\right)\left(1 - j\frac{f_k}{f}\right)}{\left(1 - j\frac{f_3}{f}\right)}$$

Then, in polar form, we can write

$$|K_L'| = |K_m| \frac{\sqrt{1 + \left(\frac{f_k}{f}\right)^2}\sqrt{1 + \left(\frac{f_4}{f}\right)^2}}{\sqrt{1 + \left(\frac{f_1}{f}\right)^2}\sqrt{1 + \left(\frac{f_3}{f}\right)^2}\sqrt{1 + \left(\frac{f_5}{f}\right)^2}}$$

$$\theta_L' = \tan^{-1}\left(\frac{f_1}{f}\right) + \tan^{-1}\left(\frac{f_3}{f}\right) + \tan^{-1}\left(\frac{f_5}{f}\right) - \tan^{-1}\left(\frac{f_k}{f}\right) - \tan^{-1}\left(\frac{f_4}{f}\right)$$

By way of summary:

$$f_k = \frac{1}{2\pi R_k C_k} \qquad f_4 = \frac{1}{2\pi R' C_f} \qquad f_5 = \frac{1}{2\pi R_f C_f}$$

$$f_1 = \frac{1}{2\pi R_g C_c} \qquad f_3 = f_k(1 + g_m R_k)$$

Application of these equations is facilitated if the amplifier is designed so that $f_k = f_1$, and $f_3 = f_4$. For this special case, the equations for voltage amplification and phase shift reduce to

$$|K_L'| = \frac{K_m}{\sqrt{1 + (f_5/f)^2}} \qquad \Theta_L' = \tan^{-1}\left(\frac{f_5}{f}\right)$$

Under these assumptions it is apparent that both the phase and frequency characteristics have been improved.

The following design procedure should be followed:

(1) R_g, C_c, R_k, R_L, and g_m are either known or evaluated from other design considerations.

(2) By assumption, $f_k = f_1$, or $R_k C_k = R_g C_c$, hence $C_k = C_c(R_g/R_k)$.

(3) Then calculate f_k and f_3 from

$$f_k = \frac{1}{2\pi R_k C_k} \qquad f_3 = f_k(1 + g_m R_k)$$

(4) R_f should be determined from power supply considerations as discussed in Art. 3.11. Now since

$$f_3 = f_4 = \frac{1}{2\pi R' C_f} = \frac{1}{2\pi(R_L R_f/R_L + R_f)C_f} = \left(\frac{1}{2\pi R_f C_f}\right)\left(\frac{R_L + R_f}{R_L}\right)$$

Then, substituting for the first bracket yields

$$f_3 = f_5\left(\frac{R_L + R_f}{R_L}\right)$$

Solving for f_5

$$f_5 = f_3\left(\frac{R_L}{R_L + R_f}\right) = \frac{1}{2\pi R_f C_f}$$

f_5 can now be calculated and C_f found from

$$C_f = \frac{1}{2\pi f_5 R_f}$$

3.14 Experimental Techniques—Measurement of Shunt Capacitance

In the example given in the preceding article, a value for the distributed wiring capacitance was assumed. Since this capacitance is normally a high percentage of the total shunt capacitance, its value cannot merely be assumed, but must be determined rather accurately. Since its value depends upon the complex geometrical configuration of the interstage coupling network, there is no practical method of calculating the wiring capacitance analytically and it is necessary to measure it experimentally.

One of the most convenient methods of measuring the total shunt capacitance (which determines the wiring capacitance since all other terms are known) is to assume a test value of plate load resistance which is of the same order of magnitude as the value of R_L which will be used. Designate this resistance as $R_L(\text{test})$.

When the basic amplifier has been constructed, this $R_L(\text{test})$ is inserted in the circuit and a gain vs. frequency characteristic is determined experimentally. The upper half-power frequency f_2 is determined from this characteristic. Then, since

$$f_2 = \frac{1}{2\pi C_s R_L(\text{test})}$$

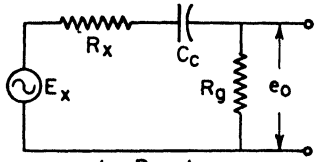
it is possible to calculate C_s directly as

$$C_s = \frac{1}{2\pi f_2 R_L(\text{test})}$$

This value of C_s, and the value of upper half-power frequency required by the design can then be used to calculate the correct value of plate load resistance.

3.15 Experimental Techniques—Square-Wave Analysis

In Art. (3.5) the equivalent circuits for a single video-amplifier stage were developed for the high- and low-frequency ranges. These circuits are reproduced in Figs. 3.30. From the point of view of conventional transient analysis, observe that the output waveform in the low-frequency equivalent circuit is the same as the transient voltage across the resistor in a series R-C circuit. However, the output voltage from the high-frequency equivalent circuit is characteristically the same as the voltage across the condenser in a series R-C circuit. Thus, assuming that the voltage $(g_m e_g)R_L$ to be a step

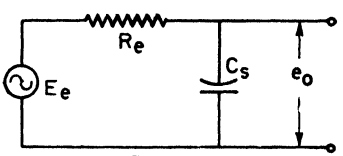


$$E_x = ue_g\left(\frac{R_L}{R_L + r_p}\right) \simeq g_m e_g R_L$$

since $R_L \ll r_p$

$$R_x = \left(\frac{R_L\, r_p}{R_L + r_p}\right) \simeq R_L$$

Low frequency equivalent plate circuit



$$E_e = ue_g\left(\frac{R_y}{R_y + r_p}\right) \approx g_m e_g R_L$$

where $R_y = \dfrac{R_g R_L}{R_g + R_L} \approx R_L$

$$R_e = \frac{r_p R_L R_g}{r_p R_L + r_p R_g + R_g R_L} \approx R_L$$

High frequency equivalent plate circuit

Fig. 3.30. Equivalent plate circuits.

function, such as would be obtained by suddenly applying a d-c voltage to the terminals of the networks, then the equations for the output voltage are:

For the low-frequency case,

$$e_0 = g_m e_g R_L\left(\frac{R_g}{R_g + R_x}\right)\epsilon^{-t/(R_x+R_g)C_c}$$

For the high-frequency case,

$$e_0 = g_m e_g R_L(1 - \epsilon^{-t/R_eC_s})$$

In both cases the initial condenser voltages were assumed to be zero. These equations may be written approximately as:

For the low-frequency case,

$$e_0 = g_m e_g R_L\, \epsilon^{-t/R_gC_c} \qquad \text{since} \qquad R_g \gg R_x$$

For the high-frequency case,

$$e_0 = g_m e_g R_L(1 - \epsilon^{-t/R_LC_s}) \qquad \text{since} \qquad R_g \gg R_L$$

During the initial period following the application of the step voltage ($g_m e_g R_L$), the high-frequency equivalent circuit controls the response. This condition arises due to the capacitive voltage divider formed by the coupling condenser C_c and the shunt capacitance C_s. Since $C_c \gg C_s$, the coupling condenser appears as a short circuit,

compared to C_s, during the initial period. Consequently, the expression for the initial voltage rise is

$$e_0 = g_m e_g R_L(1 - \epsilon^{-t/R_L C_s})$$

Thus, the voltage rises to 63.2 per cent of its maximum value in one time constant $R_L C_s$.

In a comparatively short time the shunt capacitance is charged up and then appears as an open circuit with the result that the circuit response is thereafter governed by the charging of the coupling condenser through the low-frequency equivalent plate circuit. The equation for the output voltage is then, in the approximate form,

$$e_0 = g_m e_g R_L \epsilon^{-t/R_g C_c}$$

The waveform appears as shown. The output voltage will drop 63.2 per cent below its maximum value in one time constant $R_g C_c$.

Thus, the initial rise in voltage is controlled by the charging of the shunt capacitance through the high-frequency equivalent circuit. The remainder of the waveform is controlled by the charging of the coupling condenser through the low-frequency equivalent circuit.

By replacing the step voltage with a square wave, a steady-state, or repeating transient may be obtained which can then be used with an ordinary cathode ray oscilloscope. Since

$$f_2 = \frac{1}{2\pi R_L C_s} = \text{upper half-power frequency}$$

and

$$f_1 = \frac{1}{2\pi R_g C_c} = \text{lower half-power frequency}$$

then, by graphically determining the time constants from the oscilloscope waveform, as shown in Figs. 3.31 and 3.32, the upper and lower half-power frequencies can be obtained. Furthermore, by

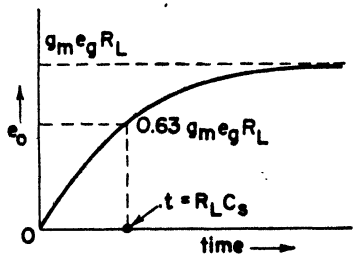


Fig. 3.31. High-frequency transient response.

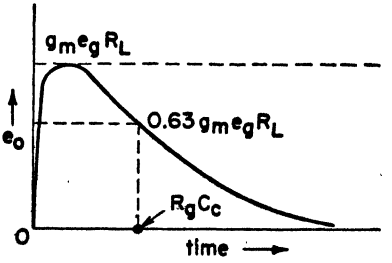


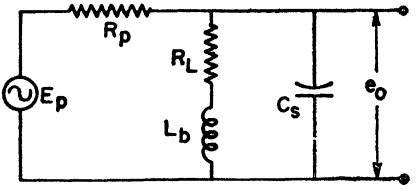
Fig. 3.32. Low-frequency transient response.

comparison of the waveforms obtained with the input square wave, it is possible to deduce the character of the distortion being introduced by the interstage coupling network.

Practically it is necessary to observe the waveform at two different square-wave frequencies, one low and one high. This procedure is required because, when the frequency is low, the half period of the square wave is very long compared to the time constant of the high-frequency equivalent circuit so that the exponential rise in output voltage occupies such a small fraction of the total waveform that it would hardly be noticeable. However, the low-frequency time constant would be readily measurable. On the other hand, when a high-frequency square wave is used, the time constant of the high-frequency equivalent circuit *is* an appreciable fraction of the square-wave half period. Thus, the exponential rise is readily visible and the time constant easily measured. However, the square-wave half period is so short compared to the low-frequency equivalent circuit time constant that the low-frequency transient hardly has time to get under way before the half period ends. Consequently, the transient charging of the coupling condenser is not observed.

When shunt peaking is used, the high-frequency equivalent plate circuit appears as shown in Fig. 3.33. The degree of compensation can be judged from the nature of the output voltage, that is, whether it is over, under, or critically damped. A similar analysis can be made of the low-frequency compensating circuit.

If any question should arise concerning the character of the distortion being introduced, the amplifier should be considered as a four-terminal network and an analysis of the type made in Art. 1.7 of Chap. 1 should be performed. This will produce the attenuation (or gain) and time delay characteristics as a function of frequency. Deductions as to the nature and comparable extent of the distortion can then be made directly from these curves.



$$E_p = ue_g\left(\frac{R_g}{R_g + r_p}\right)$$

$$R_p = \left(\frac{R_g r_p}{R_g + r_p}\right)$$

Fig. 3.33. Equivalent plate circuit when shunt peaking is used.

3.16 IF and RF Amplifiers

When an intermediate or radio frequency carrier is modulated by the frequency spectrum of a video signal, the resulting spectrum consists of the usual upper and lower sidebands centered about the carrier frequency. As a consequence, the amplifier must possess *band-pass* characteristics as opposed to the low-pass filter characteristic of the video amplifier. Proper amplification of this modulated signal requires the transmission of a comparatively wide band of frequencies, so that circuits used in this capacity are commonly called *wide-band* amplifiers. For the purposes of this text, wide-band amplifiers are considered to be those having bandwidths from 2 or 3 mcps to 20 or 30 mcps. A *high-gain* amplifier is construed to mean one which exhibits a gain in the neighborhood of 100 db.

Tubes currently in common use as intermediate frequency amplifiers in UHF superheterodyne receivers are the 6AC7 and 6AK5. Others are under development. The principal difference between *IF* and *RF* amplifiers lies in the tubes used rather than in the circuits. At the present writing, current tubes are such that *RF* amplification is feasible only at frequencies below about 1000 mcps. Type 954 and 955 (acorn) tubes are used as *RF* amplifiers up to about 300 or 400 mcps. The GL 446 Lighthouse triode is used at frequencies up to 1000 mcps.

Because *RF* amplifiers are seldom used, most of the amplification desired must be developed in the *IF* channel. The theoretical choice of the intermediate frequency is largely based upon the necessity of reducing or removing all harmonic and image responses. The *RF* preselector circuit in the receiver input will ordinarily exhibit sufficient selectivity to remove the harmonics, but the image rejection problem is more acute. Conventional preselector circuits increase their image rejection capabilities as the spacing between the true frequency and its image increases. This indicates that the image rejection is improved through the use of high intermediate frequencies. Unfortunately, the amplifier noise figure increases with increasing frequency. Moreover, at high frequencies, small changes in capacitance shift the alignment of the interstage coupling networks to a greater degree and the effects of cathode lead inductance are more pronounced. Aside from these electrical problems, the purely mechanical problem of ganging the local oscillator and pre-

selector is made more difficult when the two circuits operate at widely different frequencies.

Neglecting "practical" considerations based upon the commercial availability of components, it is evident that if the intermediate frequency is selected on the basis of image rejection, then the choice should proceed from an analysis of the *RF* preselector.

The most widely used intermediate frequency is 30 mcps and there has been a striking tendency toward standardization at this value. Although 60 mcps, and even 200 mcps, have been used, 30 mcps is gaining widespread acceptance because of the decreased signal-to-noise ratio at higher frequencies. Lower *IF*'s have been used, as low as 10 mcps.

3.17 Wide-Band Amplifier Interstage Coupling Networks

The general expression for the voltage amplification of an amplifier circuit has the form

$$|K| = g_m \left| \frac{Z_L}{1 + (Z_L/r_p)} \right|$$

where Z_L = transfer impedance of the interstage coupling network as a function of frequency. When the variational plate resistance r_p is very much greater than the transfer impedance Z_L, then the voltage amplification equation reduces to

$$|K| = g_m |Z_L|$$

It is evident that the amplifier frequency selectivity characteristic will be determined by the transfer impedance of the interstage coupling circuit. The video amplifiers discussed in the preceding articles were designed so that this impedance, and thus the amplifier gain, exhibited low-pass filter properties. In an *IF* amplifier, band pass filter action is required.

The two simplest circuits exhibiting the necessary band pass frequency characteristic are

(1) Parallel-tuned circuit (single-tuned coupling).
(2) Double-tuned circuit (transformer, or T, or π section).

Because several stages of amplification are invariably required, these two fundamental circuits, combined with the cascading of

amplifier stages, offer at least three possibilities for over-all coupling arrangements. They are

(1) N identical single-tuned stages (*synchronous tuning*).
(2) N single-tuned stages, some tuned to different frequencies (*stagger tuning*).
(3) N identical double-tuned stages.

Coupling circuits for use in wide-band amplifier service are usually compared according to

(1) Bandwidth characteristics.
(2) Selectivity characteristics.

The bandwidth is almost universally defined as the frequencies included between the two frequencies at which the power is one half of the maximum power, or between the points at which the gain is down 3 db from maximum. There are a few instances in which manufacturers have specified bandwidth in terms of half-voltage points instead of half-power points, but such cases are the exception. The only qualification to the general definition of bandwidth given is that, if the gain characteristic is "lumpy" instead of smooth, then no dip shall be more than 3 db below any peak. The principal requirement placed upon the bandwidth characteristic of the circuit used is that it be sufficiently wide to transmit the required intelligence with the desired fidelity of reproduction.

Another factor influencing bandwidth selection is the possibility of frequency drift in the local oscillator that would cause the actual intermediate frequency developed to move out of the amplifier pass band. Lastly, in some UHF systems the received signal is intermittent, due to any one of a variety of reasons. In any event, it is possible that the receiver can be tuned through the signal frequency so fast that no response is produced. The probability of this occurrence is reduced by using increasingly larger *IF* amplifier bandwidths at higher radio frequencies. The following figures are representative of actual operation.

Radio Frequency	*IF Amplifier Bandwidth*
Up to 1000 mcps	4 mcps
1000-3000 mcps	10 mcps
3000-up mcps	20 mcps

The selectivity characteristic necessary for proper operation in a

given system is chosen from a consideration of three principal factors:

(1) Bandwidth.
(2) Rejection of frequencies outside the pass band.
(3) Transient response.

Bandwidth considerations were discussed in the preceding paragraphs, principally being a problem in *passing* the desired frequencies. The *rejection* of undesired frequency components presents a difficulty in specification of terms that is not present in bandwidth definitions. It has been inconvenient to define any one term that will be suitable for general use, principally because of the vast *RF* band covered and the variety of radio services present. The word used to indicate off-band rejection is ordinarily *skirt selectivity*, a steep skirt on the selectivity characteristic giving better rejection than a gradual slope. In practice, the most convenient figure has been the *60 to 6 db bandwidth ratio*, which is the ratio of the bandwidth between the points at which the gain is down 60 db from maximum, to the bandwidth between the points at which the gain is down 6 db from maximum. A figure of 4 is generally adequate.

In steady-state operation, transient response is not a fundamental consideration. However, for services such as radar and television, the transient response is a decidedly important circuit property. Good transient response normally results from a rounded selectivity curve, which may be undesirable from skirt selectivity requirements. Flat-topped curves have a poor transient response.

The circuit elements comprising the interstage coupling network may be ordinary lumped constants of more or less conventional form. However, in some cases, particularly at the higher frequencies, the inductances in the coupling circuit are replaced by short-circuited transmission lines less than a quarter of a wavelength long. In other cases, resonant cavities are used. This technique reduces the losses, which are often greater than those needed from bandwidth considerations. In a parallel-tuned circuit, the transmission bandwidth is governed by the losses in the circuit. If the losses are very low, the bandwidth is too narrow, and the tuned circuit must be further loaded to obtain the required broad band. Generally, at ultrahigh frequencies, the losses in conventional circuits are so great that the bandwidths are larger than required, and the gain is correspondingly too low.

3.18 Synchronous Single Tuning

The essential elements of a single-tuned coupled *IF* amplifier are shown in Fig. 3.34. The capacitance, indicated by the dotted lines,

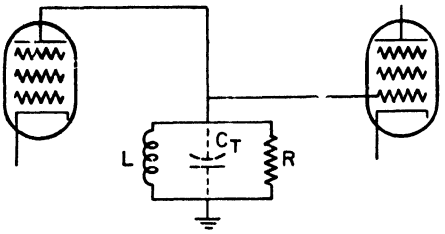


Fig. 3.34. Essential elements of a single-tuned *IF* amplifier.

that tunes the circuit to resonance is due to the tube and interstage wiring capacitances. That is,

$$C_T = C_{\text{out}}(T\text{-}1) + C_{\text{wiring}} + C_{\text{in}}(T\text{-}2)$$

According to Eq. (3.85), the gain of the stage is

$$K = g_m Z_L$$

and from Fig. 3.34 it is evident that

$$Z_L = \frac{1}{\frac{1}{R} + \frac{1}{j\omega L} - \frac{\omega C_T}{j}} = \frac{R}{1 + j\left(2\pi f R C_T - \frac{R}{2\pi f L}\right)}$$

But, at the resonant frequency $f_r = (1/2\pi)\sqrt{1/LC}$, the circuit Q is

$$Q_r = \frac{R}{2\pi f_r L} = 2\pi f_r C_T R$$

Hence, the expression for the transfer impedance becomes

$$Z_L = \frac{R}{1 + jQ_r\left(\frac{f}{f_r} - \frac{f_r}{f}\right)}$$

The complex voltage amplification is then

$$K = \frac{g_m R}{1 + jQ_r\left(\frac{f}{f_r} - \frac{f_r}{f}\right)} \tag{3.86}$$

At resonance, $f = f_r$ and $K = K_r$

so $$K_r = g_m R$$

The equation for the voltage amplification is then

$$K = \frac{K_r}{1 + jQ_r\left(\frac{f}{f_r} - \frac{f_r}{f}\right)} = \frac{K_r}{1 + jQ_r\left(\frac{f^2 - f_r^2}{ff_r}\right)}$$

The absolute magnitude of this amplification is

$$|K| = \frac{|K_r|}{\sqrt{1 + Q_r^2\left(\frac{f^2 - f_r^2}{ff_r}\right)^2}} \tag{3.87}$$

Now, since the circuit Q at resonance can be written as

$$Q_r = \frac{f_r}{\Delta f}$$

then the magnitude of the voltage amplification may be expressed in the alternative form

$$|K| = \frac{|K_r|}{\sqrt{1 + \left(\frac{f^2 - f_r^2}{f\,\Delta f}\right)^2}} \tag{3.88}$$

The half-power points at f_1 and f_2 occur when

$$Q_r\left(\frac{f^2 - f_r^2}{ff_r}\right) = \pm 1$$

Consequently, at the frequency f_1

$$Q_r\left(\frac{f_1^2 - f_r^2}{f_1 f_r}\right) = -1, \qquad \text{but} \qquad Q_r = 2\pi f_r C_T R$$

so that rearranging terms in powers of f_1 yields

$$f_1^2 + \frac{f_1}{2\pi RC_T} - f_r^2 = 0$$

Application of the standard formula for the solution of quadratic equations gives

$$f_1 = \frac{1}{2}\left(-\frac{1}{2\pi RC_T} + \sqrt{\left(\frac{1}{2\pi RC_T}\right)^2 + 4f_r^2}\right)$$

By using the same method for the upper half-power frequency f_2, the following result will be obtained:

$$f_2 = \frac{1}{2}\left(+\frac{1}{2\pi RC_T} + \sqrt{\left(\frac{1}{2\pi RC_T}\right)^2 + 4f_r^2}\right)$$

The bandwidth, $\Delta f = f_2 - f_1$, is then

$$\Delta f = \frac{1}{2\pi RC_T}$$

It is interesting to note that the bandwidth is independent of the center frequency.

The product of the voltage amplification at resonance and the bandwidth is

$$F_a = |K_r|\Delta f = \frac{g_m}{2\pi C_T}$$

which is a constant that is indicative of the merit of the stage, the symbol F_a being used to indicate the amplifier figure of merit, or gain-bandwidth product. The relationship for F_a is significant because it shows that gain and bandwidth are inseparable. An increase in amplification requires a proportional decrease in bandwidth, and conversely. Maximum performance would occur with tubes having a high mutual transconductance and low interelectrode capacitances.

When a number of identical stages are cascaded, the circuit is said to be *synchronous single-tuned.* The magnitude of voltage amplification is

$$|K_n| = \left[\frac{|K_r|}{\sqrt{1 + \left(\frac{f^2 - f_r^2}{f\,\Delta f}\right)^2}}\right]^n \qquad \text{where } n = \text{number of stages}$$

Now, let $|K_r|^n \doteq |K_r|_n$ = amplification of the cascaded n stages at resonance

Then solve for $(f^2 - f_r^2)$ obtaining

$$f^2 - f_r^2 = \pm f\,\Delta f\sqrt{\left(\frac{|K_r|_n}{|K_n|}\right)^{2/n} - 1} \tag{3.89}$$

Define terms as follows:

f_1' is the lower half-power frequency of the cascaded amplifier, f_2' is the upper half-power frequency of the cascaded amplifier, Δf, the

bandwidth of a single stage $= f_2 - f_1$, and $\Delta f'$, the bandwidth of the cascaded amplifier $= f_2' - f_1'$.

Thus, at the over-all lower half-power frequency of the cascaded amplifier, Eq. (3.89) reduces to

$$(f_1')^2 - (f_r)^2 = -f_1' \, \Delta f \sqrt{2^{1/n} - 1}$$

because $|K_n| = .707 \, |K_r|_n$ at f_1' and f_2'

Similarly, at the upper half-power frequency of the cascaded amplifier

$$(f_2')^2 - (f_r)^2 = f_2' \, \Delta f \sqrt{2^{1/n} - 1}$$

Subtracting this equation from the preceding one yields

$$(f_2')^2 - (f_1')^2 = (f_2' + f'_1) \Delta f \sqrt{2^{1/n} - 1}$$

The left-hand side of this equation may be factored as shown below.

$$(f_2')^2 - (f_1')^2 = (f_2' + f_1')(f_2' - f_1') = (f_2' + f_1') \Delta f \sqrt{2^{1/n} - 1}$$

Cancelling similar terms on either side of this equation then gives the

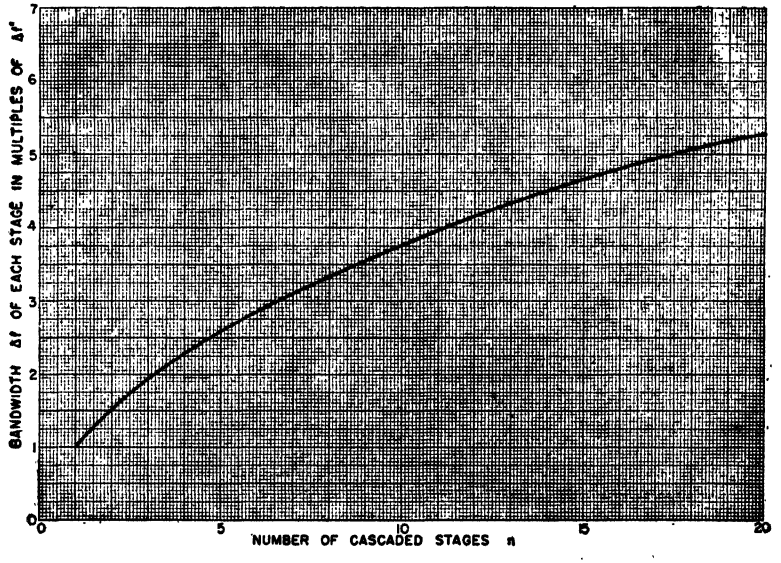


Fig. 3.35. This graph shows how the bandwidth of the individual stages must be increased, to keep the overall bandwidth constant, as the number of cascaded single-tuned stages is increased.

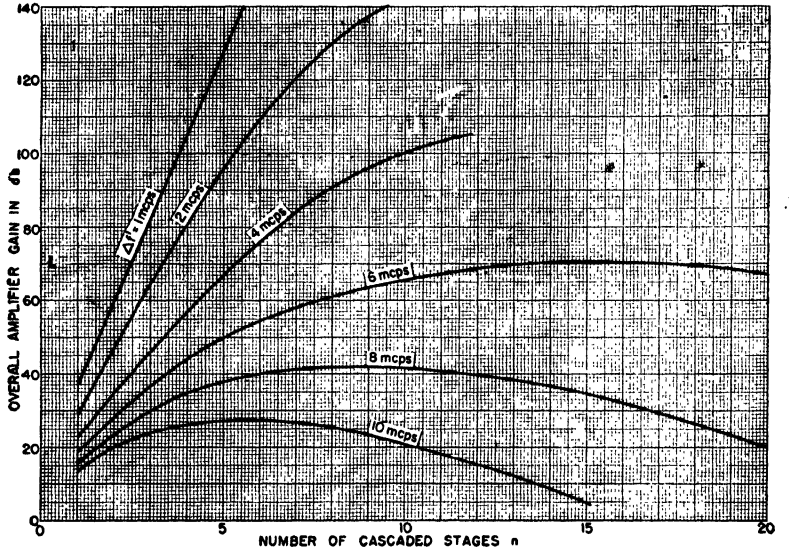


Fig. 3.36. Available voltage gain, for a given overall bandwidth, as a function of the number of cascaded single-tuned stages (for a typical pentode).

expression for the bandwidth of the cascaded amplifier in terms of the bandwidth of a single stage and the number of stages. That is,

$$\Delta f' = f_2' - f_1' = \Delta f\sqrt{2^{1/n} - 1}$$

It is evident from this equation and from Fig. 3.35 that the bandwidth of the individual stages, Δf, must increase rapidly as the number of stages is increased in order to maintain the same over-all bandwidth, $\Delta f'$. A similar relationship was developed in connection with video amplifiers and the values of the *bandwidth reduction factor*, $\sqrt{2^{1/n} - 1}$, for values of n, may be found in Table 5. Figure 3.36 presents the picture in an even more striking manner. The over-all amplifier gain has been plotted against the number of stages, with the over-all bandwidth $\Delta f'$ as a parameter, for a typical tube such as the 6AC7 or 6AK5. The curves will vary somewhat with the tube type, but their nature remains the same. Note that, for the case pictured, up to overall bandwidths of about 4 mcps, the synchronous single-tuned circuit will give high gain for a reasonable number of stages. However, for larger over-all bandwidths from about 8 mcps and beyond, it is evident that a maximum point is reached

beyond which the addition of more stages actually causes a decrease in the gain. Thus, it may or may not be possible to obtain the required amplification at the desired over-all bandwidth simply by multistaging single-tuned circuits tuned to the same center frequency.

By way of summation, the relative advantages and disadvantages of the synchronous single-tuned *IF* amplifier may be listed as follows:

(1) Advantages:
   (a) Good transient response.
   (b) Ease of adjustment.
   (c) Simplicity of design and manufacture.
   (d) Relative freedom from critical loading.

(2) Disadvantages:
   (a) Requires too many stages for over-all bandwidths in excess of about 4 mcps.
   (b) Low-skirt selectivity.

3.19 Double Tuning

When simplicity of construction and alignment and good transient response are essential design considerations, the synchronous single-tuned coupling of the preceding article may provide a satisfactory engineering solution. Additional advantages may be obtained, however, if the designer is willing to sacrifice the simplicity of the circuit. The first such step toward improvement, accompanied by complication, is the use of double-tuned circuits as the interstage coupling network in the amplifier.

The tuned transformer immediately springs to mind. Its equivalent circuit and frequency characteristics are shown in Fig. 3.37. The capacitances C_1 and C_2 are made up of the input and output capacitances of the circuits being coupled together. The circuit, when adjusted near critical coupling, gives a flat-topped response, good skirt selectivity, wider bandwidth, and higher gain. This is

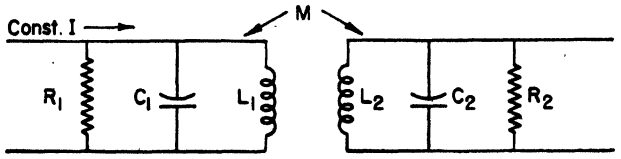


Fig. 3.37(a). Equivalent circuit of a double-tuned transformer.

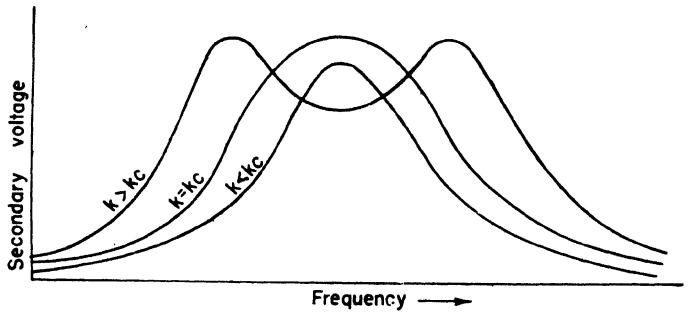


Fig. 3.37(b). Frequency characteristic of a double-tuned transformer, for different values of coupling

accompanied by an increase in the number of tuning elements, and consequently more tuning adjustments that are not independent of one another. The transient response is subject to large overshoots, especially when the coupling exceeds the critical value.

The circuit would perform effectively, but it is difficult to adjust at the frequencies normally used in UHF receivers, largely due to the delicate adjustments of the coefficient of coupling that are required. Fortunately the double-tuned transformer can be replaced by its equivalent T or π section, as indicated in Fig. 3.38, and since all of the arms of these equivalent networks are physically realizable inductances, the T or π sections can be used directly instead of the transformer. The conversion formulas, which can be obtained from any standard communication engineering text, are

$$L_a = \frac{L_1L_2 - M^2}{L_2 - M} \qquad L_A = L_1 - M$$

$$L_b = \frac{L_1L_2 - M^2}{M} \qquad L_B = L_2 - M$$

$$L_c = \frac{L_1L_2 - M^2}{L_1 - M} \qquad L_C = M$$

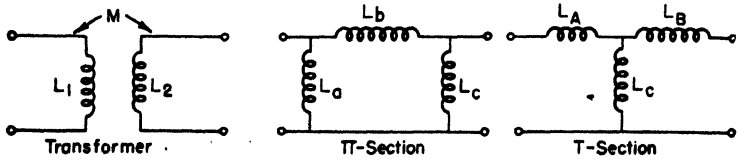


Fig. 3.38. Equivalent circuits.

If the T or π section is used, it must be recalled that the design assumes zero mutual coupling between the individual inductances and care must be exercised to minimize any tendencies toward mutual coupling. Convenience of construction requires that all three coils be wound on the same coil form, and mutual coupling is reduced by interposing a short-circuited turn between each pair of coils.

Due to the interrelationships between the three double-tuned circuits mentioned (transformer, T, π), and indicated by the preceding conversion formulas, discussion of one circuit applies equally well to the other two. Consequently, consider the tuned transformer coupling of Fig. 3.39(a). The equivalent plate circuit is given in Fig. 3.39(b). The primary capacitance C_1 is the sum of the output capacitance of tube 1, the distributed capacitance of the primary, and any added capacitance. The secondary capacitance C_2 is the sum of the input capacitance of tube 2, the distributed wiring capacitance of the secondary, and any added tuning condenser. The distributed capacitance between windings is indicated by C_m, but its effect is sufficiently small so that it may be neglected in the analysis that follows.

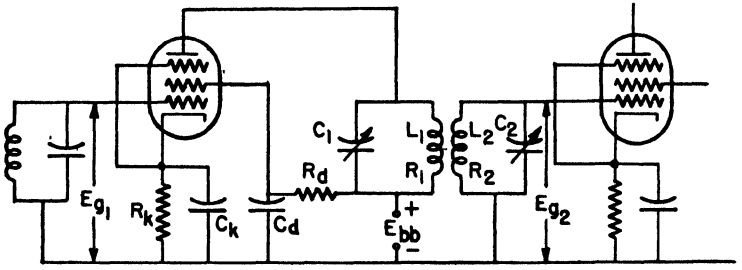


Fig. 3.39(a). Double-tuned transformer-coupled amplifier.

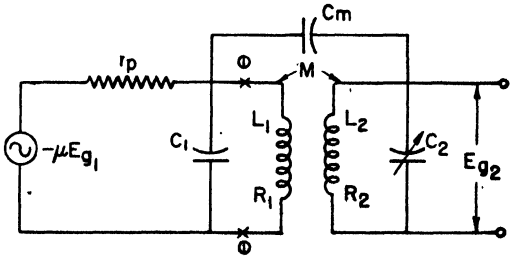


Fig. 3.39(b). Equivalent circuit of Fig. 3.39(a).

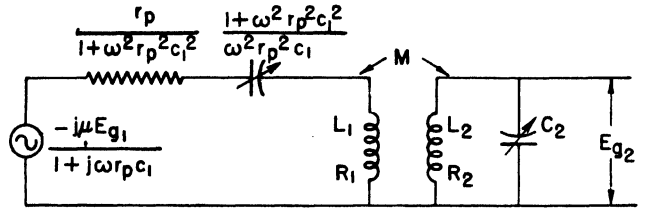


Fig. 3.39(c). Thevenized equivalent circuit of Fig. 3.39(b).

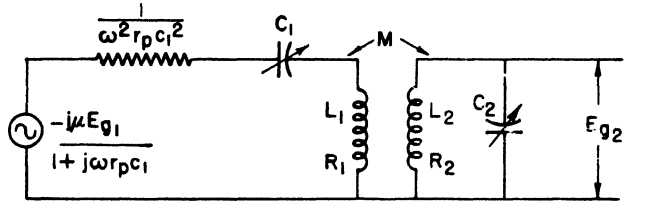


Fig. 3.39(d). Practical equivalent circuit of the double-tuned, transformer-coupled amplifier of Fig 3.39(a).

Application of Thevenin's theorem to that part of the circuit to the left of terminals (1)-(1) yields the circuit of Fig. 3.39(c). In most cases the following inequality exists.

$$\omega^2 C_1^2 r_p^2 \gg 1$$

so that the series tuning capacitor

$$\frac{1 + \omega^2 C_1^2 r_p^2}{\omega^2 C_1 r_p^2}$$

is approximately equal to C_1, giving the circuit of Fig. 3.39(d), which is the practical equivalent circuit.

The output voltage is

$$E_{g2} = -jI_2 X_{C_2} = -j\left(\frac{-jg_m E_{g1}}{\omega C_1}\right)\left(\frac{1}{Z'_{12}}\right)X_{C_2}$$

where Z'_{12} = transfer impedance $= \dfrac{Z_{11}Z_{22} - Z'_{12}}{-Z_{12}}$

The voltage amplification is then

$$K = \frac{E_{g_2}}{E_{g_1}} = -\left(\frac{g_m}{\omega^2 C_1 C_2}\right)\frac{1}{Z'_{12}}$$

Evaluate the transfer impedance by determining the mesh impedances Z_{11} and Z_{22}, and the mutual impedance Z_{12}.

$$Z_{11} = \frac{1}{\omega^2 C_1^2 r_p} + R_1 + j\left(\omega L_1 - \frac{1}{\omega C_1}\right)$$
$$Z_{22} = R_2 + j\left(\omega L_2 - \frac{1}{\omega C_2}\right)$$
$$Z_{12} = -j\omega M$$

However, at resonance, the inductive and capacitive reactances in each tuned circuit are equal. That is

$$\omega L_1 = \frac{1}{\omega C_1} \qquad \text{and} \qquad \omega L_2 = \frac{1}{\omega C_2}$$

so that the equations for mesh and mutual impedance reduce to

$$Z_{11} = \frac{1}{\omega_r^2 C_1^2 r_p + R_1}$$
$$Z_{22} = R_2$$
$$Z_{12} = -j\omega_r M$$

Thus, at resonance, the transfer impedance is

$$(Z'_{12})_r = \frac{R_2}{j\omega_r M}\left(\frac{1}{\omega_r^2 C_1^2 r_p} + R_1 + \frac{\omega_r^2 M^2}{R_2}\right)$$

Substituting into the gain equation then gives the voltage amplification at resonance as

$$K_r = j\frac{g_m \omega_r M}{\omega_r^2 C_1 C_2 R_2\left(\frac{1}{\omega_r^2 C_1^2 r_p} + R_1 + \frac{\omega_r^2 M^2}{R_2}\right)}$$

If primary and secondary are tuned to the same frequency, a condition which is most common, then

$$\omega_r^2 = \frac{1}{L_1 C_1} = \frac{1}{L_2 C_2}$$

and the primary and secondary Q's may be written directly from Fig. 3.39(d) as

$$Q'_1 = \frac{\omega r L_1}{(1/\omega_r^2 C_1^2 r_p + R_1)} \qquad Q_2 = \frac{1}{\omega_r C_2 R_2}$$

The coefficient of coupling of the transformer is

$$K = \frac{|M|}{\sqrt{L_1 L_2}}$$

Substitution of these relationships into the equation for the voltage amplification at resonance yields

$$|K_r| = \frac{g_m K \omega_r \sqrt{L_1 L_2}}{(1/Q_1' Q_2) + R^2}$$

By taking the derivative of this equation with respect to k and equating the result to zero, the value of k for maximum amplification at resonance may be obtained. Its value is

$$K_c = \frac{1}{\sqrt{Q_1' Q_2}} = \text{coefficient of } \textit{critical} \text{ coupling}$$

When k is equal to the coefficient of critical coupling, maximum power transfer from primary to secondary occurs.

In most cases the primary Q, Q_1', is less than that of the secondary. Of course, when the coupling is critical, maximum voltage amplification is obtained regardless of the relative values of the circuit Q's. However, when the Q's are unequal, the coupling may be extended *beyond* the critical point without obtaining a double-peaked response curve. If this inequality in circuit Q's exists, the coefficient of coupling for maximum amplification at resonance is called the coefficient of *optimum* coupling, and has a value

$$K_0 = \sqrt{\frac{1}{2}\left(\frac{1}{Q_1'^2} + \frac{1}{Q_2^2}\right)}$$

When optimum coupling is used, the bandwidth is larger than with critical coupling.

Because there are at least the two possible values for coupling, critical and optimum, which give maximum voltage amplification at resonance, depending upon the relative values of the circuit Q's, it is found expedient to define a more general term called *transitional* coupling* which is that value of coupling for which the transition from a single- to a double-peaked response occurs.

*Aiken, C. B., *"Two Mesh Tuned Coupled Circuit Filters,"* *Proc. IRE*, **25**, 230; 672 (1937).

The development of the expression for the bandwidth between 3 db points is extremely long, and graphical methods are generally used. It has a value

$$\Delta f' = \sqrt{2}\,\Delta f \sqrt[4]{2^{1/n} - 1}$$

The gain-bandwidth product F_a has a value

$$\sqrt{2}\left(\frac{g_m}{2\pi C_T}\right)$$

where $C_T = C_1 + C_2$ and $C_1 = C_2$. When C_1 and C_2 are not equal, the factor F_a is increased in the ratio

$$\frac{C_1 + C_2}{\sqrt{2}(C_1 C_2)}$$

an obvious advantage of some importance, relative to the single-tuned circuit, when large bandwidths at low amplification are required.

One disadvantage of double tuning that is of considerable practical significance is the difficulty of alignment. Ordinarily, it is done stage by stage and is a very tedious process. Triple tuning, while yielding a flatter response curve, wider bandwidths, and improved skirt selectivity, only increases the problems of alignment to the point where the technical feasibility of the circuit is overridden by consideration of the impracticability.

One of the obvious advantages of double tuning is the change in value of the bandwidth reduction factor. Its value, $\sqrt{2}\sqrt[4]{2^{1/n} - 1}$, is such that multistaging is accomplished more readily than in single tuning where excessive increases in the stage bandwidth were required.

3.20 Stagger Tuning

The preceding article on double-tuned coupling methods discussed the possibility of sacrificing the simplicity of single tuning in order to get an improved amplifier figure of merit and higher skirt selectivity. It would be very desirable to maintain the simplicity of the single-tuned circuit, but obtain the advantages of double tuning. The solution to this problem of circuit characteristics has been provided through the use of *stagger tuning*.

In synchronous tuning, a series of cascaded amplifiers are coupled together by single-tuned circuits resonant at the same frequency.

In stagger tuning, the cascaded tubes are connected together by single-tuned circuits having different resonant frequencies and Q's. The simplest such circuit is the *staggered pair* in which the single tuned circuits associated with two *IF* amplifier stages are tuned, by equal amounts, above and below the desired intermediate frequency. If the total separation between their resonant frequencies, which is the amount by which they are staggered, equals their individual bandwidths, then the over-all response of the two stages will be the same as that obtained from a transitionally coupled double-tuned circuit. This would yield a bandwidth that would be $\sqrt{2(2^{1/n} - 1)}$ times as large as that obtainable with synchronous tuning, but without the alignment difficulties of the double-tuned circuit.

Because the staggered pair has the same response as the transitionally coupled, double-tuned circuit, it is evident that the over-all bandwidth will go down less rapidly than for synchronous tuning. In a six-stage amplifier, the three staggered pairs would have twice the over-all bandwidth for the same gain as would six synchronous stages. The only advantage lost has been the ease of alignment, because it is now necessary to use two frequencies instead of one. This is not a severe limitation, however, because the adjustments are independent of one another and the alignment requires only two separate steps instead of one. The adjustment is slightly more critical than synchronous coupling, though certainly less than for double tuning. By all odds, stagger tuning offers the best solution to the problem of obtaining wide-band high-gain amplifiers, and the circuit is finding a wide commercial application due to its simplicity, economy, and excellent electrical characteristics.

Further improvements in the gain-bandwidth product may be made if the designer is willing to add additional frequencies for alignment purposes. Thus, by using a staggered triple, the response may be made the same as for a transitionally coupled triple-tuned circuit. The staggering process can be continued indefinitely, depending upon the extent to which loss in simplicity is justified by improved characteristics. Because each tuning adjustment is independent of all others, the sacrifice is not excessive.

3.21 Gain Limitations Imposed by Noise

Random noise in ordinary amplitude-modulated radios, and in public address systems, is common to everyone. In television and

radar, the presence of such random electrical noise produces a very undesirable interference with the visual signal. In types of data presentation depending upon intensity modulation of the electron beam, the noise appears characteristically as flashes of light occurring at random intervals and degrees of intensity. In other systems of presentation where displacement of the electron beam is used to indicate a received signal, the noise appears as a random displacement of the beam, having the appearance of a waving field of grass. In any case, the noise is highly objectionable when it becomes of the same order of magnitude as the received signal.

In general, the noise entering an amplifier (or receiver) remains substantially constant when averaged over some specified period of time. On the other hand, the received signal may, in the case of television, become quite weak when reception from a comparatively distant station takes place. In a radar set, as the target moves further and further away from the transmitter, the received echo becomes weaker and weaker. Thus, the ratio of the signal to the noise becomes smaller and smaller until the noise is an appreciable and objectionable fraction of the total received signal. Increasing the gain of the amplifier magnifies the noise as well as the desired signal so that no improvement in circuit response can be thus effected. As a result, it becomes necessary to specify the maximum sensitivity of a high-gain amplifier in terms of the signal-to-noise ratio.

3.22 Character of Noise

Since the maximum practicable sensitivity is governed by the noise component appearing in the total signal, designers of receivers and amplifiers have made extensive statistical studies of the character of noise in order to ascertain its nature and origin, and to establish some quantitative method by which it can be compared to the desired signal. The results of these studies indicate that noise may be represented by a series of small sinusoidal voltages ranging in frequency from zero to infinity and having purely random relative phase angles.

In order to measure noise quantitatively, the instantaneous components of noise current are squared, and then averaged over some specified period of time. This results in a quantity which is proportional to noise power. If noise is present due to several independent

sources, the total noise is obtained by adding the quantities obtained by the process above.

Since amplifiers are essentially bandpass devices, they only amplify those components of the noise which lie in the pass band of the amplifiers. Consequently, noise power appearing in the amplifier output will always be proportional to the bandwidth of the amplifier, thus placing another restriction upon the design. Of course, noise due to microphonics, hum in the power supply, and other disturbances do not follow this rule.

The sources of noise may be broken down into three general classes:

(1) Thermal agitation noise.
(2) Tube noises.
(3) Antenna noises.

Thermal agitation noise results from the random motion of molecules, atoms, and charges in a resistor. Were Maxwell's equations true for all microscopic phenomena, noise of this character could not occur. The eventual corpuscular nature of electricity makes the effect unavoidable. The *RMS* noise voltage in a frequency band Δf is given by

$$E_n = \sqrt{4kTR\,\Delta f}$$

where k equals $(1.38)(10^{-23})$ joules per degree Kelvin, T is the temperature in degrees Kelvin, R is the resistance in ohms, and Δf is the bandwidth in cycles per second. Since the noise energy is distributed equally throughout the frequency spectrum, the location of the amplifier bandpass Δf is of no importance.

In general, tube noises, like thermal agitation noises, are due to the discreteness of electrical charge. Variations in plate current which arise due to nonuniform emission from the cathode produce a component of noise voltage called *shot noise*. It has a distribution that is the same as thermal agitation noise.

In tetrodes and pentodes, where there is more than one positive electrode, shot effect may occur in both circuits plus a random irregular division of current between the electrodes. The latter effect causes an apparent increase in the shot noise, but is usually distinguished by the term *interception noise*.

As the electrons progress through the cathode-anode space, they induce charges on the grid. Thus, as an electron approaches the grid,

it induces a positive charge, causing it to move up through the grid leak resistor. Then as the electrons pass beyond the grid, the induced charge falls back down through the grid leak. Thus, a current is produced in the grid leak resistor, causing a voltage to be developed. If the electron beam were uniform, that is, if all the electrons traveled at the same velocity and at uniform intervals, no net voltage would be developed because the same number of electrons would be approaching the grid as were leaving. However, due to random emission, the flow of electrons is not constant and irregularities will always exist. The component of noise voltage so produced is called *induced grid noise.*

Noise entering a receiver (which is merely an amplifier which also performs frequency translation) from the antenna, is usually classified according to whether it is man-made (due to ignition systems, jamming devices, diathermy machines, etc.), or atmospheric. It is usually calculated by assuming that it is produced by thermal agitation in the radiation resistance of the antenna.

PROBLEMS

3.1 A single-stage 6AC7-pentode video amplifier is to be designed having a frequency response that is down 3 db from its mid-frequency value at 20 cps and 5 mcps. The following data apply:

$C_{\text{wiring}} = 4\mu\mu f$ $\quad r_p = 1$ megohm
$C_{cp} = 0.020\ \mu\mu f$ $\quad R_g = 500$ K
$C_{ok} = 10\ \mu\mu f$ $\quad g_m = 9000$ micromhos
$C_{pk} = 5\ \mu\mu f$

Calcualte R_L, C_c, and $|K_m|$.

3.2 A radar set operates at a pulse repetition frequency of 400 pps with a pulse width of 1 μ sec. It has been determined that satisfactory pulse reproduction occurs if two loops of the Fourier series spectrum coefficient are passed by the amplifier. Design a one-stage video amplifier, uncompensated, capable of passing the necessary frequencies. Use a 6AC7 tube and assume the wiring capacitance in the plate circuit to be $5\mu\mu f$. If the amplifier works directly into the intensity grid of a CRT for which $C_{\text{in}} = 10\mu\mu f$, find R_L and C_c.

3.3 Repeat Problem 3.2 if the pulse width is changed to a half of a microsecond at a 20 pps repetition rate.

3.4 A single-stage video amplifier is connected between the second detector and intensity grid of a CRT in a radar set. The CRT intensity grid

input capacitance is 10$\mu\mu$f. The amplifier is to have a bandwidth ($\pm$3 db) from 20 cps to 4 mcps. A 6AG7-type tube is to be used for which

$C_{gp} = 0.06\ \mu\mu f$	$r_p = 100$ K
$C_{out} = 7.5\ \mu\mu f$	$g_m = 7700$ micromhos
$C_{in} = 12.5\ \mu\mu f$	$R_g = 500$ K
$E_b = 300$ volts	$I_b = 28$ ma
$E_{sc} = 300$ volts	$I_{sc} = 7$ ma
$E_c = -2$ volts	

After the amplifier has been built, a load of 6000 ohms is placed in the plate circuit and the upper half-power frequency is found experimentally to be 1 mcps.

(a) Calculate E_{bb}, R_L, R_d, R_k, C_k, $|K_m|$

(b) Calculate the value of shunt-peaking inductances required for optimum phase and optimum frequency responses.

(c) Calculate the constants of a low-frequency correction circuit, assuming $E'_{bb} = 550$ volts. Use optimum compensation. Recalculate R_d.

(d) Calculate the amplification and phase angle at 20 cps and 4 mcps, with and without compensation, to determine the extent of the improvement in each case.

3.5 A video amplifier to be used in a television receiver is to be designed to meet the following minimum specifications:

(a) over-all bandwidth = 20 cps to 5 mcps ($\pm$3 db).

(b) over-all minimum gain = 55 db.

Use as many stages of amplification as required, but the preamplifiers are to be 6AC7's while the final amplifier is to use a 6AG7. Assume 8$\mu\mu$f distributed capacitance in each interstage network.

(a) Determine the number of stages required, and their individual bandwidths, assuming no compensation.

(b) Determine the values of all circuit elements required.

(c) Calculate and plot the amplitude and phase characteristic of the entire amplifier.

3.6 Redesign the amplifier of Problem 3.5 using high-frequency compensation ($N = 0.5$) in each stage *now* required. Calculate the constants of a low-frequency compensating circuit.

3.7 A six-stage synchronous single-tuned *IF* amplifier has an over-all voltage gain of 120 db at the intermediate frequency of 30 mcps. The tubes used have

$$g_m = 5000 \text{ micromhos} \qquad C_o = 5\mu\mu f \qquad C_i = 11\mu\mu f$$

The interstage wiring capacitance is 14$\mu\mu$f in each case.

(a) Find the value of load resistance in the coupling circuits.

(b) Calculate the over-all bandwidth.

(c) What is the value of L in the tuned circuits?

(d) If economic considerations were such that only 4 stages were permissible, what load resistance would be required for an over-all gain of 120 db?

(e) What would be the over-all bandwidth of the 4-stage amplifier?

3.8 A typical triode for which

g_m = 2000 micromhos r_p = 10,000 ohms R_k = 600 ohms

operates as an amplifier without a cathode by-pass condenser. Calculate the effective transconductance, plate resistance, and amplification factor under this condition.

3.9 If the resistor of Problem 3.8 is by-passed by a condenser C_k, what are the effective values of g_m, r_p, and μ at the frequency for which the reactance of C_k equals R_k?

REFERENCES

Books

Brainerd, J. G., Reich, H. J., Koehler, G., and Woodruff, L. F., *Ultra-High-Frequency Techniques*, Van Nostrand, 1942.

Eastman, A. V., *Fundamentals of Vacuum Tubes*, McGraw-Hill, 1941.

Everitt, W. L., *Communication Engineering*, McGraw-Hill.

Fink, D. G., *Principles of Television Engineering*, McGraw-Hill, 1940.
Radar Engineering, McGraw-Hill, 1947.

M.I.T. Staff, *Principles of Radar*, McGraw-Hill, 1946.
Applied Electronics, Wiley, 1943.

Vacuum Tube Amplifiers, Rad. Lab. Tech. Series, Vol. 18, McGraw-Hill.
Microwave Receivers, Vol. 23, McGraw-Hill.

Radio Research Laboratory Staff, *Very High Frequency Techniques*, Mc-Graw-Hill, 1947.

Sarbacher, R. I., and Edson, W. A., *Hyper and Ultrahigh Engineering*, Wiley, 1943.

Strutt, M. J. O., *Ultra- and Extreme-Short Wave Reception*, Van Nostrand, 1947.

Wilson, J. C., *Television Engineering*, Pitman.

Periodicals

Barber, "Video Amplifier Design," *Communications*, 18 (6), 13 (1938).

Barco, "Measurement of Phase Shift in Television Amplifiers," *RCA Review*, Apr. 1939.

Baum, "Design of Broad Band I. F. Amplifiers," *Applied Phys.*, **17** 519 (1949).

Bedford, A. V., and Fredendall, G. L., "Transient Response of Multistage Video Amplifiers," *Proc. IRE*, Apr. 1939.

Builder, G., "Amplification of Transients," *Wireless Eng.*, May 1935.

Butterworth, S., "On the Theory of Filter Amplifiers," *Wireless Eng.*, **7**, 536 (1930).

Carson, J. R., "Fluctuation Noise in Vacuum Tubes," *Bell System Tech. J.*, Oct. 1934.

Donely and Epstein, "Low Frequency Characteristics of Coupling Networks," *RCA Review*, Apr. 1942.

Everest, "Wideband Television Amplifiers," *Electronics*, Jan. and May 1938.

Freeman, R. L., and Schantz, D. D., "Video Amplifier Design," *Electronics*, Aug. 1937.

Herold, E. W., "High Frequency Correction in Resistance Coupled Amplifiers," *Communications*, Aug. 1938.

Janski, "Natural Static," *Proc. IRE*, Jan. 1939.

Johnan, and Llewellyn, F. B., "Limitations to Amplification," *Elec. Eng.*, Nov. 1934.

Keall, O. E., "Correction Circuits for Amplifiers," *Marconi Rev.*, May 1935.

Lane, H., "Resistance-Capacitance Amplifier in Television," *Proc. IRE*, Apr. 1932.

Llewellyn, F. B., "A Study of Noise in Vacuum Tubes and Attached Circuits," *Proc. IRE*, Feb. 1930.

Lynch, "Video Amplifier for Low Frequency Correction," *Communications*, Apr. 1943.

Nagy, "Design of Vision Frequency Amplifiers," *Television*, Mar., Apr., and May 1937.

North, D. O., and Ferris, W. R., "Fluctuations Induced in Vacuum Tube Grids at High Frequency," *Proc. IRE*, Feb. 1941.

Nyquist, H., "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.*, July 1928.

Oakey, "Distortionless Amplification of Electrical Transients," *Wireless Eng.*, May 1931.

Petrich, "A Wideband 550 mcps Amplifier," *Proc. IRE*, **35**, 1371 (1947).

Preissman, A., "Some Notes on Video Amplifier Design," *RCA Review*, Apr. 1938.

Preissman, A., "Low Frequency Square Wave Analysis," *Communications*, Mar. 1942.

Pollack, D., "Choice of Tubes for Wideband Amplifiers," *Electronics*, Apr. 1939.

Puckle, O. S., "Transient Aspect of Wideband Amplifiers," *Wireless Eng.*, May. 1935.

Seeley, S. W., and Kimball, C. N., "Analysis and Design of Video Amplifiers," *RCA Review*, Jan. 1939.

Wheeler, H. A., "Wideband Amplifiers for Television," *Proc. IRE*, July 1939.

Swift, "Amplifier Testing by Means of Square Waves," *Communications*, Feb. 1939.

CHAPTER 4
INTRODUCTION TO TRANSMISSION LINES

THE subject of transmission lines is ordinarily introduced in courses preliminary to the one for which this text is intended. However, an effective transition from this generalized theory to the specific applications treated in Chap. 5 necessitates a concise recapitulation of the basic phenomena. Consequently, Chap. 4 is of the nature of a review and can, in some cases, be omitted if warranted by student preparation.

4.1 Principles of Wave Propagation

Before analyzing electrical wave motion, consider a water analogy. Assume that, by some means, a small volume of water in a pool is set into vertical oscillatory motion. However, due to the elasticity and density of the water, this small volume cannot oscillate independently of the water immediately adjacent to it. As a result, the excesses and deficiencies of pressure are transmitted to the surrounding water, which receives energy and is also put into motion. This volume of water does the same thing to the water shell surrounding it, and as the procedure repeats, a water wave is propagated across the pool. The wave could have been directed if such a disturbance had been established in a narrow canal, the canal walls serving to guide the motion of the strain.

The propagation of this wave represents the transmission of energy from the initial volume of water across the pool, or down the canal, yet there is no net displacement of the water molecules in the direction of energy transmission. They are displaced vertically during the incidence of the wave, and they resume their quiescent position after the wave has passed. The velocity of the wave represents the velocity of energy transmission, not the velocity of the water molecules.

Considering now the case of an indefinitely long transmission line being supplied by an a-c generator, it is apparent that a similar situation exists. At the terminals of the generator the dielectric is being strained in an oscillatory manner. Such a strain represents an oscillatory displacement of the dielectric and such a displacement cannot occur independently of the medium surrounding that portion under strain. Hence, the strain, or displacement, is transmitted to the adjoining medium, and so on, continuously as guided by the conductors. The wires guide the motion of the disturbance since they alone have the "free" charges necessary to terminate the lines of electric field intensity.

Since the strain is oscillatory in time, in the steady state, there is no net or progressive movement of any one line of electric field intensity. As in the case of the water wave, only the strain progresses. However, due to the oscillatory characteristic of the electric field, it does *appear* to move in the same way that the water in a water wave *apparently* (but not actually) progresses.

Because of this mechanism of wave motion, it is evident that the energy being transmitted is merely guided by the conductors, actually being present in the dielectric medium between and around the conductors. Broadly speaking then, a transmission line is really a mechanism for guiding electrical energy through the dielectric. Any analysis of transmission line phenomena must proceed from a recognition of this fact. The presence of the dielectric is acknowledged by the character of the transmission line parameters of series inductance and shunt capacitance and conductance. The presence of imperfect metallic conductors adds a series resistance term.

4.2 Types of Transmission Lines

Radio frequency transmission lines are generally of three basic types:

(1) Two-wire line.
(2) Four-wire line.
(3) Coaxial line.

The physical conformations of these lines are shown in Fig. 4.1. The two-wire line consists of two parallel conductors separated by a distance which is small compared to the wavelength of the signal on the line. For this reason it is used principally in the lower RF regions, being replaced by coaxial line or waveguide in the microwaves. If

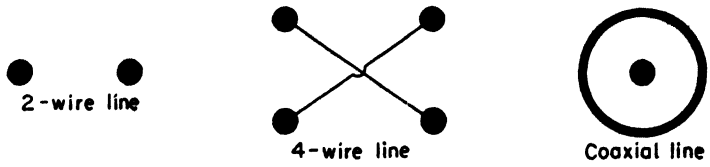


Fig. 4.1. Types of RF transmission lines.

the line is very far away from other objects or circuits, or symmetrically located with respect to them, the currents in both conductors are the same.

The four-wire line consists of four parallel wires, with diagonally opposite wires connected in parallel. Because of this feature, the four-wire line has less external field than the two-wire line and is less affected by neighboring circuits.

The coaxial line consists of two conductors, arranged coaxially with respect to one another as shown in Fig. 4.1. The center conductor is held in place by several methods ranging from the use of solid dielectric material inside the cable to the use of small insulating beads, or in some cases, stub supports. It is used to the exclusion of open-wire lines at elevated frequencies up to about 3000 mcps, and in some cases at 10,000 mcps.

4.3 Losses in Transmission Lines

The losses present in transmission lines usually arise from three separable and distinct sources:

(1) Losses in the conductor material (copper loss).
(2) Losses in the insulating material (dielectric loss).
(3) Losses due to coupling into other circuits either by radiation or by magnetic induction.

Generally speaking, all three types of losses increase with frequency.

The increase in copper loss with frequency is primarily due to skin effect. The effective depth of penetration δ of the current into the conductor is a function of frequency according to the relation

$$\delta = \sqrt{\frac{\rho}{\pi \mu f}}$$

where ρ is the resistivity in meter-ohms, μ, the permeability in henries

per meter, and f = frequency in cycles per second. This effective depth of current penetration determines the effective cross-sectional area A of conductor through which conduction occurs. Then since the resistance is given by

$$R = \rho \frac{L}{A}$$

and the area is evaluated according to

$$A = \pi[r^2 - (r - \delta)^2]$$

where r is the radius of the conductor, then

$$R = \text{constant } \sqrt{\text{frequency}}$$

because δ depends upon frequency. As the frequency increases, the skin depth decreases, decreasing the area A, and therefore increasing both the resistance R and the power loss.

The dielectric loss arises primarily from leakage currents in the insulating material. In most cases this loss is quite small. For air, it is negligible. Solid dielectrics, which are frequently used in coaxial lines, seldom have such desirably low conductivities. As a consequence, resort is made to support of the inner conductor by small, solid, dielectric beads. The major part of the dielectric thus remains air, and the losses are kept to a minimum. However, as the frequency is increased, the bead spacing becomes an appreciable fraction of a wavelength and the dielectric discontinuities caused by the beads produce reflections which lower the overall efficiency. Another possibility exploited at frequencies above 3000 mcps is to support the inner conductor with stubs, a tactic which is extremely effective unless the frequency of the signal on the line is somewhat unstable, causing variations in the electrical-line length.

The induction losses arise from coupling, through the magnetic field, to other nearby circuits. The coupling takes place in the same manner as in transformers. The effect of the circuits thus coupled into the line is to insert an effective resistance in series with the distributed resistance, effectively increasing the total loss. In any case, since the energy is removed from the system and is not delivered to the load, it is a definite loss. Since the coaxial line is almost completely self-shielding, it is not as subject to this effect as is the open-wire line.

Even though the magnetic field is not directly coupled into other circuits, some coupling to free space is bound to occur, resulting in radiation of energy. Again, the effect of the medium into which radiation occurs is to present the radiating system with an effective resistance, measuring the amount of energy lost from the system. This resistance increases the over-all line losses, furthering the drop in efficiency. This loss varies approximately as the square of the frequency, and inversely as the spacing between conductors in the open-wire line. Radiation losses in coaxial lines are extremely small because of the shielding action of the outer conductor.

4.4 Derivation of the Telegraphers' Equation

The derivation of useful equations for transmission line calculations is somewhat complicated by the fact that the circuit constants of resistance, inductance, and capacitance are not lumped in the sense ordinarily associated with circuit elements. Instead, they tend to be uniformly distributed throughout the length of the transmission line. Consequently, ordinary circuit theory cannot be applied directly to long line problems. A further difference exists because wave motion takes place on transmission lines as opposed to the ordinary stationary oscillation found in the usual lumped-constant circuit treated by a-c theory. Wave motion occurs because the physical length of the line is large compared to the wavelength of the signal. In any case, it is an appreciable fraction of the wavelength of the disturbance.

In its most general form, any transmission line consists of two conductors separated by a dielectric medium. Associated with each of the conductors is a series resistance, most conveniently expressed in ohms per unit length and designated by R. Due to the magnetic fields associated with the currents on the line, the inductance per unit length (L) is also uniformly distributed, and in series with the resistance R. Assuming a practical dielectric, then in addition to the capacitance per unit length (C) produced by the geometry of the transmission line and the nature of the dielectric, there is also a leakage conductance (G). Considering any differential length of the line, it could be represented by the four-terminal network shown in Fig. 4.2. The complete transmission line could then be represented by an infinite number of these sections connected in cascade, as the differential section approaches an infinitesimally small length.

The change in voltage with respect to the distance l measured along the line can then be written directly from Fig. 4.2 as

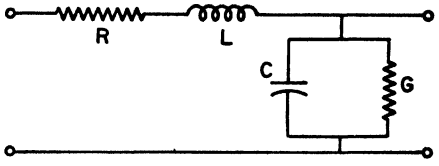


Fig. 4.2. Differential length of transmission line.

$$\frac{\partial v}{\partial l} = Ri + L\frac{\partial i}{\partial t} \tag{4.1}$$

where v is the instantaneous voltage and i, the instantaneous current.

In the same manner, the change in current with respect to distance along the line may be written as

$$\frac{\partial i}{\partial l} = Gv + C\frac{\partial v}{\partial t} \tag{4.2}$$

Differentiate Eq. (4.1) with respect to l, obtaining

$$\frac{\partial^2 v}{\partial l^2} = R\frac{\partial i}{\partial l} + L\frac{\partial}{\partial t}\left(\frac{\partial i}{\partial l}\right) \tag{4.3}$$

Now substitute Eq. (4.2) into Eq. (4.3).

$$\frac{\partial^2 v}{\partial l^2} = R\left(Gv + C\frac{\partial v}{\partial t}\right) + L\frac{\partial}{\partial t}\left(Gv + C\frac{\partial v}{\partial t}\right)$$

Carry out the indicated multiplication and differentiation. Then, collect terms to obtain

$$\frac{\partial^2 v}{\partial l^2} = (RG)v + (RC + LG)\frac{\partial v}{\partial t} + (LC)\frac{\partial^2 v}{\partial t^2} \tag{4.5}$$

By following exactly the same procedure, but starting with the differentiation of Eq. (4.2), an identically similar expression can be obtained for the current. That is

$$\frac{\partial^2 i}{\partial l^2} = (RG)i + (RC + LG)\frac{\partial i}{\partial t} + (LC)\frac{\partial^2 i}{\partial t^2} \tag{4.6}$$

These equations are the *Telegraphers' equations*. They are the fundamental differential equations which must be satisfied if wave motion is to occur on the transmission line. Consequently, they are merely special forms of the *wave equation*, which is common to all fields of physics.

It is customary to assume that both the current and voltage are harmonic functions of time. The assumption is a valid one to make

because this will ordinarily be the case in practical transmission systems. However, even if the signal is not harmonic, as long as it is periodic, it can be resolved into a series of related sine and cosine terms by Fourier Series analysis. Hence, for the purposes at hand, assume that the current and voltage are harmonic functions of time as given by the following relations:

$$v = Ve^{j\omega t} \tag{4.7}$$

$$i = Ie^{j\omega t} \tag{4.8}$$

Substitute Eq. (4.7) into Eq. (4.5). Carry out the required derivatives and cancel out the exponential term on both sides of the equation. This yields

$$\frac{d^2V}{dl^2} = (RG)V + j\omega(RC + LG)V - (\omega^2 LC)V$$

Fortunately, the right hand side of this relation can be factored as follows:

$$\frac{d^2V}{dl^2} = (R + j\omega L)(G + j\omega C)V \tag{4.9}$$

This form is especially convenient since

$$Z = \text{series impedance per unit length} = R + j\omega L$$
$$Y = \text{shunt admittance per unit length} = G + j\omega C$$

Consequently, the following equation results:

$$\frac{d^2V}{dl^2} - (ZY)V = 0 \tag{4.10}$$

By the same process it can be shown that

$$\frac{d^2I}{dl^2} - (ZY)I = 0 \tag{4.11}$$

Equations (4.10) and (4.11) are merely compact forms of the Telegraphers' equations when the signal on the line is a harmonic function of time.

4.5 Solution of the Telegraphers' Equations

The Telegraphers' equations given in (4.10) and (4.11) are ordinary, second order, linear, differential equations with constant

coefficients. The solutions of such equations are standard and have the following form:

$$V = Ae^{\gamma l} + Be^{-\gamma l} \tag{4.12}$$

$$I = Ce^{\gamma l} + De^{-\gamma l} \tag{4.13}$$

where $$\gamma = \sqrt{ZY} \tag{4.14}$$

The four unknown constants A, B, C, and D must be evaluated to obtain the complete solution in practical form. This can be done in the following manner. According to Eq. (4.1)

$$\frac{\partial v}{\partial l} = Ri + L\frac{\partial i}{\partial t}$$

However, v and i were assumed to be harmonic functions of time as specified by Eqs. (4.7) and (4.8). Putting the expressions for the harmonic time dependence into Eq. (4.1) written above yields

$$\frac{dV}{dl} = (R + j\omega L)I = ZI \tag{4.15}$$

Now differentiate Eq. (4.12) to obtain

$$\frac{dV}{dl} = A\gamma e^{\gamma l} - B\gamma e^{-\gamma l} \tag{4.16}$$

Solving Eqs. (4.16) and (4.15) simultaneously for I, and replacing the constant γ by $\sqrt{ZY}$, yields

$$I = \frac{A}{\sqrt{Z/Y}} e^{\gamma l} - \frac{B}{\sqrt{Z/Y}} e^{-\gamma l} \tag{4.17}$$

The quantity in the denominator of the coefficients has the dimensions of impedance, so, for the sake of simplicity in writing the equations, let

$$\sqrt{Z/Y} = Z_c \tag{4.18}$$

Therefore, the solutions to the Telegraphers' equations may be written in the following form:

$$V = Ae^{\gamma l} + Be^{-\gamma l} \tag{4.19}$$

$$I = \frac{A}{Z_c} e^{\gamma l} - \frac{B}{Z_c} e^{-\gamma l} \tag{4.20}$$

Before completing the solutions for the unknown constants, it is necessary to establish a convention for the measurement of l, the

distance along the line. Two possibilities suggest themselves: the positive l direction may be assumed to be measured

(1) From the generator toward the load.
(2) From the load toward the generator.

The latter designation is most generally followed and will be used in this text.

Assuming the load to be specified by an impedance Z_R, the physical setup of the transmission system appears as shown in Fig. 4.3. The establishment of this convention permits the evaluation of the two remaining unknown constants, A and B, which remain in Eqs. (4.19) and (42.0). At the load end of the line,

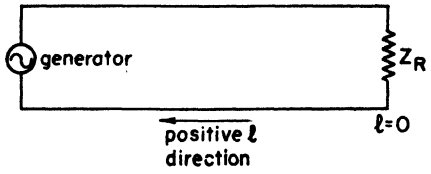


Fig. 4.3. Convention for the measurement of I.

$$\begin{aligned} l &= 0 \\ E_R &= \text{receiving-end voltage} = V \\ I_R &= \text{receiving-end current} = I \end{aligned}$$

Substitution of these terminal conditions into Eqs. (4.19) and (4.20) yields

$$\begin{aligned} E_R &= A + B \\ I_R &= \frac{1}{Z_c}(A - B) \end{aligned}$$

Solving these two equations simultaneously for A and B produces

$$A = \tfrac{1}{2}(E_R + I_R Z_c) \tag{4.21}$$

$$B = \tfrac{1}{2}(E_R - I_R Z_c) \tag{4.22}$$

Substituting these expressions for A and B back into Eqs. (4.19) and (4.20) yields the complete solution of the Telegraphers' equations in the following form:

$$V = \frac{1}{2}(E_R + I_R Z_c)e^{\gamma l} + \frac{1}{2}(E_R - I_R Z_c)e^{-\gamma l} \tag{4.23}$$

$$I = \frac{1}{2}\left(I_R + \frac{E_R}{Z_c}\right)e^{\gamma l} + \frac{1}{2}\left(I_R - \frac{E_R}{Z_c}\right)e^{-\gamma l} \tag{4.24}$$

These are called the *long-line equations.* They may be put into an alternative form, which is often more convenient for computations, by collecting terms in the following manner:

$$V = E_R\left(\frac{e^{\gamma l} + e^{-\gamma l}}{2}\right) + I_R Z_c\left(\frac{e^{\gamma l} - e^{-\gamma l}}{2}\right)$$

$$I = I_R\left(\frac{e^{\gamma l} + e^{-\gamma l}}{2}\right) + \frac{E_R}{Z_c}\left(\frac{e^{\gamma l} - e^{-\gamma l}}{2}\right)$$

These are the exponential forms of

$$V = E_R \cosh \gamma l + I_R Z_c \sinh \gamma l \tag{4.25}$$

$$I = I_R \cosh \gamma l + \frac{E_R}{Z_c} \sinh \gamma l \tag{4.26}$$

This form is generally of considerable use in design work since values of hyperbolic functions are ordinarily available in tabular form.

4.6 The Propagation Constant

The quantity γ which appears in the foregoing equations indicates the dependence of the voltage and current upon the distance measured from the load end of the line. In other words, it establishes the nature of the variations in voltage and current along the axis of propagation and is accordingly, and appropriately, called the *propagation constant.*

It was previously shown that

$$\gamma = \sqrt{ZY} \tag{4.27}$$

where Z is the series impedance per unit length and Y, the series admittance per unit length. Since both Z and Y are generally complex numbers, then it follows that, in the most general case, the propagation constant is likewise complex, having both real and imaginary terms. Hence, let

$$\gamma = \alpha + j\beta \tag{4.28}$$

where α is the attenuation constant and β, the phase (or wavelength) constant.

The attenuation and phase constants are generally complicated functions of the distributed parameters of the line and the frequency. Since

$$\gamma = \alpha + j\beta = \sqrt{ZY}$$

then

$$\alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)}$$

Square both sides and collect real and imaginary terms on either side of the equation, obtaining

$$\alpha^2\beta^2 = RG - \omega^2 LC$$
$$2\alpha\beta = \omega(RC + GL)$$

If these two equations are solved simultaneously for either α or β, a quartic equation of the following form results:

$$ax^4 + bx^2 + c = 0$$

The standard solution of this equation has the form of the square root of the solution of a quadratic equation, or

$$x = \pm\sqrt{\frac{1}{2a}[-b \pm \sqrt{b^2 - 4ac}]}$$

In both cases where the $\pm$ signs occur, they are replaced simply by $+$ signs. This is necessary because both α and β are *real positive* numbers, and such a result can be obtained only by this choice of signs. Omitting the details of the manipulation, the simultaneous solution of the two preceding equations yields

$$\alpha = \sqrt{\tfrac{1}{2}(RG - \omega^2 LC) + \tfrac{1}{2}\sqrt{(R^2 + \omega^2 L^2)(G^2 + \omega^2 C^2)}} \qquad (4.29)$$

$$\beta = \sqrt{\tfrac{1}{2}(\omega^2 LC - RG) + \tfrac{1}{2}\sqrt{(R^2 + \omega^2 L^2)(G^2 + \omega^2 C^2)}} \qquad (4.30)$$

These expressions are far too cumbersome for convenient use and certain simplifying approximations are usually invoked to bring them down into more compact form. Consider the expression for attenuation. Carry out the multiplication indicated under the second radical. This yields

$$\alpha = \sqrt{\tfrac{1}{2}(RG - \omega^2 LC) + \tfrac{1}{2}\sqrt{(RG)^2 + \omega^2(R^2C^2 + G^2L^2) + (\omega^2 LC)^2}}$$

In most practical applications the series resistance and shunt conductance are very small in comparison to the series-reactance and shunt-susceptance terms, respectively. In any case, their product is nearly always negligible in comparison to other terms in the equation. Under this assumption, the product RG may be omitted and the expression for the attenuation becomes

$$\alpha = \sqrt{-\frac{\omega^2 LC}{2} + \frac{1}{2}\sqrt{(\omega^2 LC)^2 + \omega^2(R^2C^2 + G^2L^2)}}$$

Factor this equation into the following form

$$\alpha = \sqrt{-\frac{\omega^2 LC}{2} + \frac{\omega^2 LC}{2}\sqrt{1 + \left[\left(\frac{R}{\omega L}\right)^2 + \left(\frac{G}{\omega C}\right)^2\right]}}$$

The second radical has the form of $\sqrt{(1 + x)}$ where

$$x = \left(\frac{R}{\omega L}\right)^2 + \left(\frac{G}{\omega C}\right)^2$$

When $x \ll 1$, it can be shown that this is approximately equal to

$$1 + x/2$$

The approximation is valid for most communication transmission lines because, as previously mentioned, ordinarily

$$R \ll \omega L$$
$$G \ll \omega C$$

Using this approximation, the attenuation reduces to

$$\alpha \doteq \sqrt{-\frac{\omega^2 LC}{2} + \frac{\omega^2 LC}{2} + \frac{\omega^2 LC}{4}\left[\left(\frac{R}{\omega L}\right)^2 + \left(\frac{G}{\omega C}\right)^2\right]}$$

Cancelling like terms yields

$$\alpha \doteq \frac{\omega\sqrt{LC}}{2}\sqrt{\left(\frac{R}{\omega L}\right)^2 + \left(\frac{G}{\omega C}\right)^2}$$

Complete the square of the expression under the radical, obtaining

$$\alpha \doteq \frac{\omega\sqrt{LC}}{2}\sqrt{\left(\frac{R}{\omega L} + \frac{G}{\omega C}\right)^2 - \frac{2RG}{\omega^2 LC}}$$

But, as previously noted, $RG \doteq 0$, so

$$\alpha \doteq \frac{\omega\sqrt{LC}}{2}\left(\frac{R}{\omega L} + \frac{G}{\omega C}\right)$$

which may be written in the alternative form

$$\alpha \doteq \frac{R}{2}\sqrt{\frac{C}{L}} + \frac{G}{2}\sqrt{\frac{L}{C}} \tag{4.31}$$

By following a similar process it can be shown that the original equation for the phase constant reduces to the following approximation:

$$\beta \doteq \omega\sqrt{LC} \tag{4.32}$$

The attenuation α accounts for the decrease in amplitude as the wave propagates down the line. The phase constant determines the amount that the phase of the wave is retarded as it moves in the direction of propagation. Consequently, in the direction of energy transmission, the attenuation causes the wave amplitude to diminish, and the phase constant causes the relative phase to be retarded.

4.7 Interpretation of the Solution

Each of the long-line equations is composed of two terms. According to Eq. (4.17)

$$V = \tfrac{1}{2}(E_R + I_R Z_c)e^{\gamma l} + \tfrac{1}{2}(E_R - I_R Z_c)e^{-\gamma l}$$

Considering each component separately, let

$$E^+ = \tfrac{1}{2}(E_R + I_R Z_c)e^{\gamma l} = E_R{}^+ e^{\alpha l}e^{j\beta l}$$

$$E^- = \tfrac{1}{2}(E_R - I_R Z_c)e^{-\gamma l} = E_R{}^- e^{-\alpha l}e^{-j\beta l}$$

Hence, $V = E^+ + E^-$

Similarly, for the currents

$$I = I^+ + I^-$$

where

$$I^+ = \frac{1}{2}\left(I_R + \frac{E_R}{Z_c}\right)e^{\gamma l} = I_R{}^+ e^{\alpha l}e^{j\beta l}$$

$$I^- = \frac{1}{2}\left(I_R - \frac{E_R}{Z_c}\right)e^{-\gamma l} = I_R{}^- e^{-\alpha l}e^{-j\beta l}$$

The attenuation factors $e^{\alpha l}$ and $e^{-\alpha l}$, are pure numerics affecting only the magnitudes of E^+, E^-, I^+, and I^-. On the other hand, the phase terms $e^{\pm j\beta l}$ may be written as

$$e^{\pm j\beta l} = \cos \beta l \pm j \sin \beta l = 1\underline{\pm \beta l^0}$$

Consequently, these factors affect only the phase of the wave, having no effect upon the magnitude.

The positive l direction is measured *away* from the load end. Therefore, as the point of observation of the wave is moved away from the load, toward the generator, l is positive and increasing. This causes E^+ and I^+ to increase in magnitude, whereas E^- and I^- decrease in amplitude. Moreover, E^+ and I^+ are advanced in phase (positive angle), while E^- and I^- are retarded (negative angle). Thus E^+ and I^+ have the characteristics of a wave traveling toward the load from the generator, while E^- and I^- have the properties of a wave traveling away from the load. These conclusions are based

upon the observation previously made that the amplitude of a wave decreases, and its phase is retarded, as it propagates. In the case of the E^+ and I^+ waves, they diminish in amplitude and are retarded in phase when l is negative, measured *toward* the load. On the other hand, E^- and I^- diminish and have their phase retarded when l is positive, measured *away* from the load. Thus, in the general case, we can conclude that there are four waves on a transmission line; *incident* (upon the load) current and voltage waves (E^+ and I^+), and *reflected* (from the load) current and voltage waves (E^- and I^-). The reflected wave is produced whenever there is an impedance discontinuity between the line and load. In this respect the line is similar to a water canal which is partially blocked at one end. A water wave traveling down the canal will strike this obstruction. Part of the energy in the wave is reflected and travels back along the canal in the opposite direction. The part which passes through the partial obstacle represents energy removed from the system and is comparable to the energy delivered to the load terminating the transmission line.

4.8 Characteristic Impedance

The quantity Zc appearing in the foregoing equations is called the *characteristic* (or surge) impedance. It is defined by the relationship previously given

$$Z_c = \sqrt{\frac{Z}{Y}} = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{4.35}$$

The characteristic impedance is a unique quantity. For example, assume that the transmission line is infinitely long. Under this assumption, the long-line equations

$$V = \frac{1}{2}(E_R + I_R Z_c)e^{\gamma l} + \frac{1}{2}(E_R - I_R Z_c)e^{-\gamma l}$$

$$I = \frac{1}{2}\left(I_R + \frac{E_R}{Z_c}\right)e^{\gamma l} + \frac{1}{2}\left(I_R - \frac{E_R}{Z_c}\right)e^{-\gamma l}$$

reduce to the following form:

$$V = \frac{1}{2}(E_R + I_R Z_c)e^{\gamma l}$$

$$I = \frac{1}{2}\left(I_R + \frac{E_R}{Z_c}\right)e^{\gamma l} \qquad \text{since } e^{-\infty} = 0$$

Note that there is only one exponential term, or wave, in each equation, indicating that the reflected wave has been eliminated. This is not surprising since any possible discontinuity was eliminated by the assumption. The input impedance of the infinite line is then

$$Z = \frac{V}{I} = \frac{\frac{1}{2}(E_R + I_R Z_c)e^{\gamma l}}{\frac{1}{2}\left(I_R + \frac{E_R}{Z_c}\right)e^{\gamma l}} = Z_c$$

which is the characteristic impedance. Thus, it should be noted that the infinite line has

(1) An input impedance equal to the characteristic impedance,
(2) No reflected waves.

4.9 Line Terminated in Its Characteristic Impedance

Consider an infinitely long line. Its input impedance equals the characteristic impedance. If any finite length of the line is cut off, the input impedance of the remainder is still equal to the characteristic impedance because its infinite length is unaffected by the removal of any finite section. The system then appears as shown in Fig. 4.4.

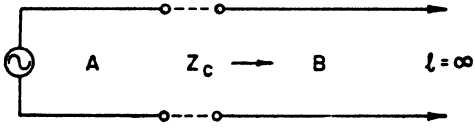


Fig. 4.4.

Section A is finite, section B is infinite. The input impedance of B is the characteristic impedance Z_c. Therefore, it could be replaced simply by a lumped impedance equal to Z_c, and the generator would be unable to distinguish the change. Hence, a line terminated in its characteristic impedance is electrically equivalent to an infinite line. This being the case, we can then deduce that there are no reflected waves on such a line and that the input impedance equals the characteristic impedance. This can be proved very quickly from the long-line equations. For the general case it was shown that

$$V = \frac{1}{2}(E_R + I_R Z_c)e^{\gamma l} + \frac{1}{2}(E_R - I_R Z_c)e^{-\gamma l}$$

$$I = \frac{1}{2}\left(I_R + \frac{E_R}{Z_c}\right)e^{\gamma l} + \frac{1}{2}\left(I_R - \frac{E_R}{Z_c}\right)e^{-\gamma l}$$

When the line is terminated in its characteristic impedance, then

$$Z_R = Z_c \qquad \text{so} \qquad I_R Z_c = E_R \qquad \text{and} \qquad \frac{E_R}{Z_c} = I_R$$

Therefore, the second terms in both equations are zero, indicating the absence of reflected waves. Furthermore, the input impedance is

$$Z = \frac{V}{I} = \frac{E_R + I_R Z_c}{I_R + \dfrac{E_R}{Z_c}} = Z_c$$

which is the characteristic impedance.

4.10 Wavelength, Phase Velocity, Time Delay

The wavelength λ of the wave on the line is the distance which it must travel for its phase to be retarded by 2π radians. Hence,

$$\beta\lambda = 2\pi$$

or

$$\lambda = \frac{2\pi}{\beta} \tag{4.36}$$

The phase of the wave changes by 2π radians in a length of time τ equal to the period of the wave. Since λ is the distance which it travels in shifting its phase by 2π radians, then the ratio of wavelength λ to the period τ is the phase velocity v_p. That is,

$$v_p = \frac{\lambda}{\tau} = f\lambda$$

$$v_p = f\left(\frac{2\pi}{\beta}\right) = \frac{\omega}{\beta} \tag{4.37}$$

The time delay is the time required for an equiphase point to travel length 'd' of the line. That is

$$\tau_d = \frac{d}{v_p} = \left(\frac{\beta}{\omega}\right)d = \frac{d}{f\lambda} \tag{4.38}$$

4.11 Dissipationless (Distortionless) Lines

For the larger part of the transmission lines used in the ultra-high frequencies the line losses are negligible, although this condition may not exist in the lower RF regions. For a lossless line, both the

series resistance and the shunt conductance are zero, and the expression for the propagation constant becomes

$$\gamma = \sqrt{ZY} = \sqrt{(R + j\omega L)(G + j\omega C} = \sqrt{(j\omega L)(j\omega C)}$$

or $$\gamma = 0 + j\omega\sqrt{LC}$$

Then, equating reals and imaginaries on either side of the equation,

$$\alpha = 0$$
$$\beta = \omega\sqrt{LC}$$

It will be recalled that this meets the requirements for no distortion as previously outlined in Chap. 3 since the attenuation is independent of frequency while the phase constant is a linear function of frequency. Consequently, the lossless line is commonly referred to as a distortionless line. For most UHF applications the approximations are quite good.

Under lossless conditions, the characteristic impedance Z_c is

$$Z_c = \sqrt{\frac{Z}{Y}} = \sqrt{\frac{L}{C}}$$

which is a pure resistance independent of frequency. The expression for the wavelength

$$\lambda = \frac{2\pi}{\beta}$$

reduces to the following form.

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

Replacing γ by $j\beta l$ in the hyperbolic form of the long-line equations puts them in the following form:

$$V = E_R \cos \beta l + jI_R Z_c \sin \beta l$$

$$I = I_R \cos \beta l + j\frac{E_R}{Z_c} \sin \beta l$$

because

$$\cosh (j\beta l) = \cos \beta l \quad \text{and} \quad \sinh (j\beta l) = j \sin \beta l$$

The input impedance at any point on the line, looking toward the load, is simply the ratio of the voltage to the current, or

$$Z = \frac{V}{I} = \frac{E_R \cos \beta l + jI_R Z_c \sin \beta l}{I_R \cos \beta l + j(E_R/Z_c) \sin \beta l}$$

Factor out a Z_c and divide numerator and denominator by cos βl. This yields

$$Z = Z_c \frac{E_R + jI_R Z_c \tan \beta l}{I_R Z_c + jE_R \tan \beta l}$$

Divide numerator and denominator by I_R, obtaining

$$Z = Z_c\left(\frac{Z_R + jZ_c \tan \beta l}{Z_c + jZ_R \tan \beta l}\right) \quad \text{where} \quad Z_R = \frac{E_R}{I_R} \tag{4.39}$$

Other useful equations may be derived by replacing the line by its input impedance given by the equation above so that the complete system has the equivalent circuit shown in Fig. 4.5. Therefore,

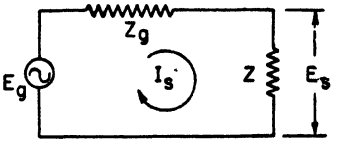


Fig. 4.5. Equivalent circuit of a transmission line at the sending end.

$$E_S = E_g\left(\frac{Z}{Z_g + Z}\right) \tag{4.40}$$

$$I_S = \frac{E_g}{Z_g + Z} \tag{4.41}$$

4.12 The Reflection Coefficient

The solution of the Telegraphers' equations indicated the presence of two current, and two voltage, waves, one of each traveling toward the load, and one of each traveling away from the load. The ratio of the reflected wave (E^- or I^-) to the incident wave (E^+ or I^+) at the load, is called the *reflection coefficient* Γ. So

$$\Gamma = \left.\frac{E^-}{E^+}\right|_{l=0} = \frac{E_R - I_R Z_c}{E_R + I_R Z_c}$$

Divide through numerator and denominator by I_R to obtain

$$\Gamma = \frac{Z_R - Z_c}{Z_R + Z_c} = \frac{(Z_R/Z_c) - 1}{(Z_R/Z_c) + 1} \tag{4.42}$$

Taking the ratio of I^- to I^+ yields

$$-\Gamma = \frac{I^-}{I^+}\bigg|_{l=0} = \frac{I_R - (E_R/Z_c)}{I_R + (E_R/Z_c)} = \frac{Z_c - Z_R}{Z_c + Z_R} = -\left(\frac{Z_R - Z_c}{Z_R + Z_c}\right) \tag{4.43}$$

Hence, when the voltage is reflected without change of phase, the current *is* reflected with change of sign, and conversely. This indicates that the phase angle between E^+ and I^+ is 180° different from that between E^- and I^-.

If the load impedance is set equal to the characteristic impedance of the line, that is, if $Z_R = Z_c$, then Γ is zero. No reflection occurs. If the load impedance is made equal to a short circuit, then $Z_R = 0$, and $\Gamma = -1 = 1\,\underline{|180°}$. Hence, $E^- = E^+\Gamma = -E^+$, but $I^- = I^+$. If the load impedance is an open circuit, $Z_R = \infty$ and $\Gamma = 1$. Thus, $E^- = E^+\Gamma = E^+$, while $I^- = -I^+\Gamma = -I^+$.

In the short-circuited case the voltage is reflected with a 180° phase reversal, but without change of magnitude, and is always in phase opposition to the incident wave. Hence, at the load, the voltage is zero. On the other hand, the current is reflected without phase reversal and is in phase addition with the incident current wave. The exact converse occurs in the open-circuited case.

In general, it is found expedient to express the reflection coefficient in polar form, as follows, in order to indicate the magnitude and phase of the reflected wave relative to the incident wave:

$$\Gamma = \left|\frac{Z_R - Z_c}{Z_R + Z_c}\right| \underline{|\phi} \tag{4.44}$$

This is quite straightforward mathematically, but the physical reasons for these results warrant additional comment. On the average, the total energy in the electromagnetic wave in the dielectric of a dissipationless line is equally divided between the electric and magnetic fields. When the wave arrives at an open-circuit termination, the magnetic field must disappear since no current can flow in an ideal open circuit. The energy stored in this field cannot simply evaporate, it must reappear somewhere. Moreover, since the presence of this open circuit requires the magnetic field to undergo a change, then this changing magnetic field produces an electric field in accordance with Maxwell's equation

$$\text{Curl } \mathbf{E} = \mu \frac{\partial \mathbf{H}}{\partial t}$$

This is basically the principle upon which transformers operate and is probably more familiar in the specialized form

$$e = -N \frac{d\phi}{dt}$$

Consequently, the energy stored in the magnetic field is converted into energy in the electric field, and this field adds vectorially to the incident electric field. This process increases the voltage at the open circuit, and reduces the current to zero. The increase in voltage starts another (reflected) wave traveling away from the load. Therefore, the voltage is reflected without change of phase. However, the reflected current wave must cancel the incident wave at the termination in order to reduce the current to zero. Thus, the reflected current wave has its phase shifted by 180° relative to the incident current wave. Since no energy is absorbed in the idealized termination, the reflected waves have the same amplitude as the incident waves.

The same reasoning process can be applied to the short-circuited case, or to any intermediate condition. Analysis from this standpoint facilitates the understanding of transmission line phenomena considerably.

4.13 Vector Representation of Traveling Waves

In the usual a-c theory courses one becomes accustomed to representing stationary oscillations in terms of rotating vectors. It is of considerable advantage in numerous applications to employ the same technique with traveling waves. Consider the incident voltage wave E^+,

$$E^+ = E_R{}^+ e^{\alpha l} e^{j\beta l}$$

or

$$E^+ = E_R{}^+ e^{\alpha l} \underline{|\beta l}$$

Assuming that the attenuation is zero, then

$$E^+ = E_R{}^+ \underline{|\beta l}$$

E^+ could be represented as shown in Fig. 4.6 and has a locus which is a circle as l increases from zero. Note that rotation is counterclockwise in the usual manner for positive angles.

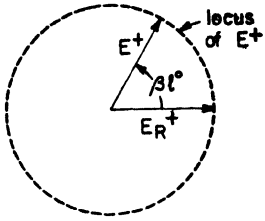


Fig. 4.6. Vector representation of the incident traveling wave (no attenuation).

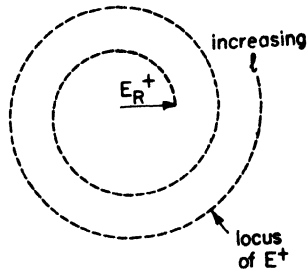


Fig. 4.7. Effect of attenuation on the locus of E^+.

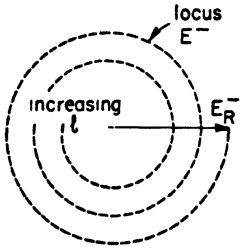


Fig. 4.8. Locus of the reflected voltage wave, showing the effect of attenuation.

If the attenuation were not zero, then the magnitude of E^+ would increase as l became larger and larger, that is, as the point of observation approaches the generator. This effect causes the locus to appear as a spiral as shown in Fig. 4.7. The same sort of representation may be made of E^-, I^+, and I^-, but proper attention must be paid to the sign of the phase angle since that determines the direction of rotation. For example, the corresponding locus for the reflected voltage wave E^- would appear as shown in Fig. 4.8, where the angle is measured clockwise, and the attenuation causes the wave amplitude to diminish as the generator is approached.

To obtain the complete vector representation of the voltage long-line equations, it would be necessary to consider the sum of these two rotating vectors, taken point by point, as indicated in

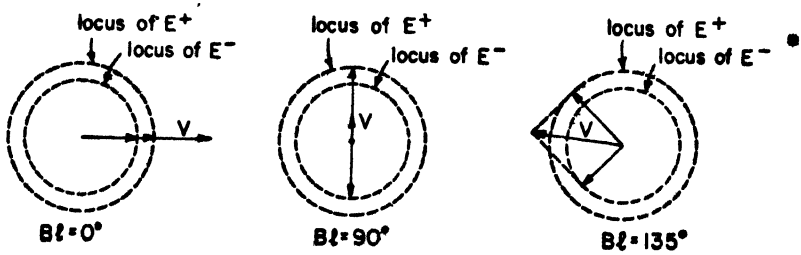


Fig. 4.9. Vector positions at different points on a lossless line.

Fig. 4.9. In this figure it has been assumed, for the sake of convenience of representation, that the attenuation is zero.

If this resultant voltage is plotted as a function of the distance l, it appears substantially as shown in Fig. 4.10. Since this wave is fixed in space, while varying with respect to time, it is called a *standing wave* to distinguish it from the *traveling wave* that is delivering energy to the load. Therefore, it is a stationary oscillation of voltage or current, as the case may be.

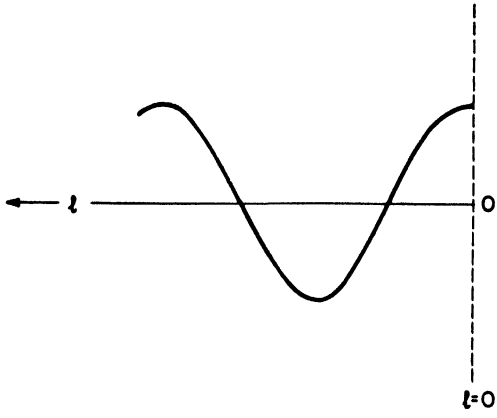


Fig. 4.10. Voltage standing wave.

4.14 Standing Waves

The reflection coefficient was shown to be

$$\Gamma = \frac{Z_R - Z_c}{Z_R + Z_c}$$

When the transmission line is short-circuited at the load end, the reflection coefficient becomes

$$\Gamma = -1 = 1 \underline{|180^\circ}$$

In other words, reflection occurs without loss in amplitude, but the reflected voltage wave is 180° out of phase with the incident wave.

Generally, the voltage across an ideal short circuit must be zero, although the current is finite in cases involving transmission lines. Thus, in constructing the vector diagrams it is advantageous to use

the current in the short circuit as the reference vector. The current at any point on the line is related to this load current by the long-line equation

$$I = \frac{1}{2}\left(I_R + \frac{E_R}{Z_c}\right)e^{\gamma l} + \frac{1}{2}\left(I_R - \frac{E_R}{Z_c}\right)e^{-\gamma l}$$

But, in the short-circuited case, E_R is zero, and this reduces the current equation to

$$I = \frac{1}{2} I_R(e^{\gamma l} + e^{-\gamma l})$$

Since l is zero at the load, then the incident and reflected current

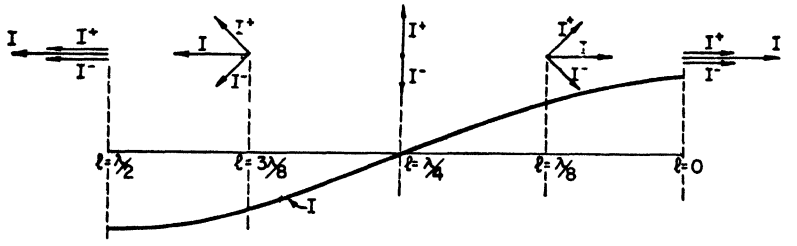


Fig. 4.11. Formation of current standing wave on a short circuited line.

waves add up quite properly to I_R, the load current. As l increases I^+ (incident wave) rotates counterclockwise, and I^- (reflected wave) rotates clockwise. Each vector rotates 2π radians for each wavelength of the line. Hence, the vector diagrams of Fig. 4.11 can be drawn. The net sum of I^+ and I^- at various points along the line, when plotted, produce the current standing wave. It is generally more expedient to plot the magnitude and phase of the standing wave

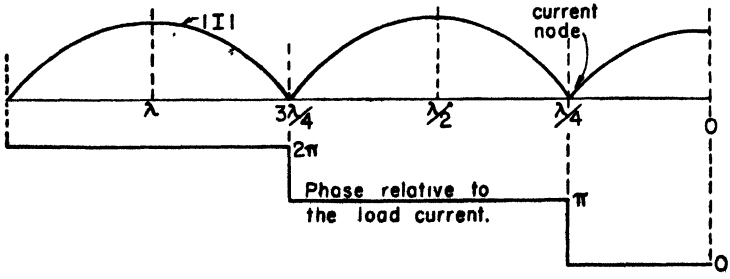


Fig. 4.12. Current standing wave on a short-circuited dissipationless line.

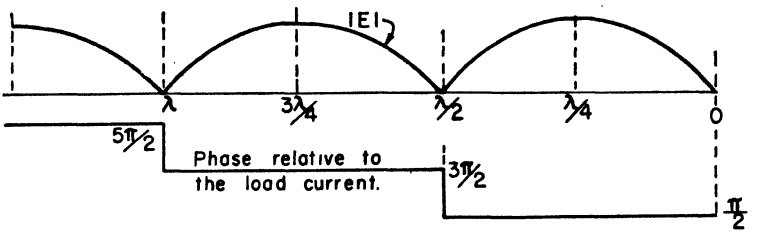


Fig. 4.13. Voltage standing wave on a short-circuited dissipationless line.

separately because practical standing-wave measuring equipment indicates magnitude only. Following this procedure, the standing wave of Fig. 4.11 appears as shown in Fig. 4.12.

The same technique may be followed to determine the voltage standing wave. The results appear as shown in Fig. 4.13. Both standing waves have been plotted in Fig. 4.14 and the scales have been chosen so that $|I_{\max}|$ and $|E_{\max}|$ are the same. The long-line equation for voltage is

$$V = \tfrac{1}{2}(E_R + I_R Z_c)e^{\gamma l} + \tfrac{1}{2}(E_R - I_R Z_c)e^{-\gamma l}$$

For the short-circuited case, $E_R = 0$, so that

$$V = \tfrac{1}{2} I_R Z_c (e^{\gamma l} - e^{-\gamma l})$$

Consequently, the magnitudes of reflected and incident waves are related according to

$$|E^+| = |E^-| = \tfrac{1}{2} I_R Z_c$$

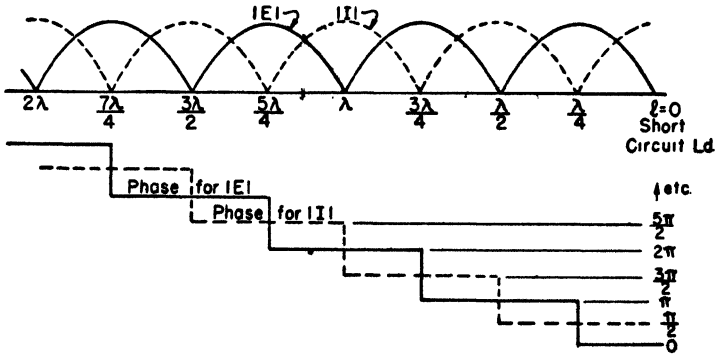


Fig. 4.14. Standing waves on a dissipationless, short-circuited line. Phase angles are measured relative to the current in the short circuit.

In the same manner for the current, it can be shown that

$$|I^+| = |I^-| = \tfrac{1}{2} I_R$$

Therefore,

$$|E^+| = |I^+|Z_c \qquad \text{and} \qquad |E^-| = |I^-|Z_c$$

Hence, the current in Fig. 4.14 has been plotted to a scale Z_c larger than the voltage scale.

Note that whenever the voltage wave is a maximum, the current wave is a minimum, and conversely, duplicating the conditions at the load every half wavelength.

If the line is opened up at any point d distance from the short circuit, the input impedance is simply the ratio of V to I evaluated at that point. From Fig. 4.14 it will be noted that whenever d is less than an odd multiple of a quarter wavelength, but greater than any even multiple, the current *lags* the voltage by 90°. This is characteristic of an inductance. On the other hand, when d is greater than any odd multiple of a quarter wavelength, but less than any even multiple, then the current *leads* the voltage by 90°, as in a capacitance. When d is exactly an odd quarter wavelength from the short, the current is zero and the input impedance is infinite, as in a parallel-tuned circuit, at resonance. When d is an even multiple of a quarter wavelength the voltage is zero and the input impedance is also zero. This corresponds to a series resonant circuit at the resonant frequency. Because of these properties, transmission lines which have standing waves are called *resonant lines*. Lines without standing waves are called *nonresonant* or *flat* lines.

For the short-circuited case, assuming dissipationless lines, the input impedance is

$$Z = \frac{V}{I} = Z_c \frac{Z_R + jZ_c \tan \beta l}{Z_c + jZ_R \tan \beta l} = \left(\frac{jZ_c \tan \beta l}{Z_c}\right) Z_c$$

or

$$Z = jZ_c \tan \beta l$$

This indicates mathematically the impedance properties discussed qualitatively above. That is, depending upon the length of the transmission line, relative to a quarter wavelength, the input impedance may be a pure inductive or capacitive reactance. Or, it may exhibit resonant or antiresonant properties. Consequently, resonant lines find a wide application at the higher radio frequencies

as circuit parameters. Although the open-circuited line has the same general capabilities, the short-circuited line is most commonly used in practical systems because it is easier to obtain an adjustable shorted line than an adjustable open line. It also possesses a higher degree of mechanical rigidity, an important consideration in many applications.

4.15 Standing-Wave Ratio

The ratio of the maximum value of the standing wave to its minimum value is defined as the *standing-wave ratio* (SWR). Hence,

$$\rho = \frac{|E_{\max}|}{|E_{\min}|} = \frac{|I_{\max}|}{|I_{\min}|}$$

Or, rewriting in terms of incident and reflected waves,

$$\rho = \frac{|E^+| + |E^-|}{|E^+| - |E^-|} = \frac{|I^+| + |I^-|}{|I^+| - |I^-|}$$

On a dissipationless line, the magnitude of the incident and reflected waves is everywhere the same. Therefore,

$$|E^+| = |E_R{}^+| \qquad \text{and} \qquad |E^-| = |E_R{}^-|$$

Substituting into the equation for the standing wave ratio yields

$$\rho = \frac{|E_R{}^+| + |E_R{}^-|}{|E_R{}^+| - |E_R{}^-|} = \frac{1 + \dfrac{|E_R{}^-|}{|E_R{}^+|}}{1 - \dfrac{|E_R{}^-|}{|E_R{}^+|}}$$

However, from the equation for the reflection coefficient

$$|\Gamma| = \frac{|E_R{}^-|}{|E_R{}^+|} = \frac{|Z_R - Z_c|}{|Z_R + Z_c|}$$

Which, when substituted into the previous equation yields

$$\rho = \frac{1 + |\Gamma|}{1 - |\Gamma|} = \frac{1 + \dfrac{|Z_R - Z_c|}{|Z_R + Z_c|}}{1 - \dfrac{|Z_R - Z_c|}{|Z_R + Z_c|}}$$

Reduction to a common denominator produces the following form:

$$\rho = \frac{|Z_R + Z_c| + |Z_R - Z_c|}{|Z_R + Z_c| - |Z_R - Z_c|} \tag{4.45}$$

It is evident then, that the standing-wave ratio may be determined experimentally by actual measurement of the standing-wave pattern, or by calculation based upon knowledge of the values of load impedance and characteristic impedance.

4.16 Impedance at a Voltage or a Current Maximum

At a current or a voltage maximum the voltage and the current are in phase as an examination of Fig. 4.14 will reveal. Consequently, if the line were opened up at such a point, the input impedance would be a pure resistance. Now, at a voltage maximum,

$$R_{in} = \frac{V}{I} = \frac{|E_{max}|}{|I_{min}|}$$

However, in the preceding article, the standing-wave ratio was defined as

$$\rho = \frac{|E_{max}|}{|E_{min}|} \qquad \text{or} \qquad |E_{max}| = \rho |E_{min}|$$

Consequently, the input impedance is

$$R_{in} = \rho \frac{|E_{min}|}{|I_{min}|} = \rho Z_c \tag{4.46}$$

Conversely, at a current maximum,

$$R'_{in} = \frac{|E_{min}|}{|I_{max}|} = \frac{|E_{min}|}{\rho |I_{min}|}$$

$$R'_{in} = \frac{Z_c}{\rho} \tag{4.47}$$

Since the characteristic impedance of a lossless line is a pure resistance, and since the standing-wave ratio is a pure numeric, it is evident in both cases that the input impedance is purely resistive.

4.17 Practical Importance of a Low Standing-Wave Ratio

Transmission lines employed in UHF communication systems may fulfill either of two general functions:

(1) To convey energy from one point to another in the system, or
(2) To act as circuit elements.

Lines used for energy transmission are operated with a low standing-wave ratio, while resonant lines (large SWR) are used as circuit

elements. A low standing-wave ratio is desired for energy transmission because:

(1) The power-handling capacity increases with decreasing SWR.
(2) The line efficiency is higher.
(3) The input impedance of the line is not as sensitive to slight changes in frequency, thus keeping the power input and output less dependent upon frequency.
(4) The danger of oscillator "pulling" is reduced.

The maximum power that can be transmitted on a line is

$$P_{\max} = \frac{E^2_{bd}}{Z_c} \qquad \text{where} \qquad E_{bd} = \text{insulation breakdown voltage}$$

When standing waves are present on the line, the voltage has higher maximum values than a corresponding flat line supplied by the same generator. Consequently, the insulation breakdown voltage is reached at a lower generator voltage than when there are no standing waves. Hence, a low SWR allows a larger power-handling capacity.

For lines a quarter of a wavelength long, or longer, the attenuation causes a reduction in efficiency when the SWR is large because the energy in the reflected wave is dissipated in heat, either directly, or after being reflected at the generator, and so on, if a mismatch exists at both ends of the line. For lines that are electrically short, such as power transmission lines, efficiency may actually be greater when the load impedance is unequal to the characteristic impedance, depending upon the value of generator internal impedance. Of course, for a lossless line the efficiency is 100 per cent regardless of the standing-wave ratio. In any case, the loss in efficiency is not appreciable, even on practical lines, except for very large standing-wave ratios.

The frequency of an oscillator is determined by the constants in its resonant circuit. Due to coupling between the oscillator and the transmission line which carries power to the antenna, an additional impedance is placed in this tank circuit which tends to affect the oscillator frequency. Transmission lines with large SWR have large input impedance variations for slight changes in frequency. This effect causes the impedance inserted in the resonant circuit to vary, "pulling" the oscillator away from its desired natural frequency. Low standing-wave ratios minimize this effect.

PROBLEMS

4.1 The characteristics of an open wire line are calculated by the following formulas:

$$R = \frac{8.3}{a}\sqrt{f}\ \mu\text{ohms/m} \qquad f = \text{frequency in cps}$$

$$L = 0.921 \log_{10} D/a\ \mu\text{hy/m} \qquad a = \text{wire radius in cm}$$

$$C = \frac{12.06}{\log_{10} D/a} \qquad D = \text{distance between wire centers in cm}$$

If $D = 2$ cm and $a = 0.04$ cm, and if $\omega = 10^9$ rad/sec, find the line constants.

4.2 If $Z_c = 276 \log_{10} D/a$ ohms, calculate (a) the characteristic impedance Z_c, (b) the attenuation constant α, (c) the phase constant β, and (d) the Q of the line given in Problem 4.1.

4.3 The line of Problem 4.2 is 5 m long and terminated in a load equal to the characteristic impedance. A generator whose open circuit voltage is 50 volts and whose internal impedance is $31 + j0$ ohms is connected at the sending end. What is the magnitude and angle of the current in, and the voltage across, the load?

4.4 The line of Problem 4.3 is short-circuited by a bar at the load end. Considering the line to be dissipationless, and supplied by the same generator, calculate the input current and short-circuit current. Sketch the current and voltage distribution along the total length of line. Show the value of $|I_{max}|$.

4.5 What is the length, in inches, of a quarter wavelength section of a transmission line operated at a frequency of 300 mcps?

4.6 A dissipationless transmission line whose characteristic impedance is 200 ohms is connected to a load of $100 + j0$ ohms. The frequency is 300 mcps. What is the input impedance of the line if the length is (a) 15 cm, (b) 25 cm, (c) 50 cm, and (d) 65 cm? What conclusion can be drawn from the similarity of some of the answers?

4.7 A 200-ohm dissipationless line is terminated in a 100-ohm resistor. If the line is one wavelength long, and if the voltage across the load is 500 volts, sketch the voltage and current distribution along the line. Calculate maximum and minimum values of voltage and current.

4.8 What is the required length of 70-ohm coaxial line to provide an inductive reactance of 250 ohms when the wavelength of the signal on the line is 80 cm?

4.9 A lossless line with a characteristic impedance of $469 + j0$ ohms is 30 cm long and operated at a frequency of $\omega = 10^9$ rad/sec.

(a) What inductance should be connected at the load end of the line to make the input impedance infinite?

(b) What capacitance should be connected at the load end to make the input impedance zero?

(c) Sketch the current and voltage distributions on the line for each of these two cases.

(d) What conclusions can you draw from the problem as to the effect of terminating reactances upon the *apparent* length of the line?

4.10 A transmission line is to be used to carry energy from the transmitter of a system to the antenna. For most effective operation of the line, should the antenna impedance be equal to the characteristic impedance of the line, or should it be the complex conjugate of the characteristic impedance? Explain your answer fully.

4.11 Find the load impedance of a lossless transmission line if $Z_c = 50$ ohms, $Z_s = 15 - j45$ ohms, and $\beta s = 70°$.

REFERENCES

Brainerd, J. G., Koehler, G., Reich, H. J., Woodruff, L., *Ultrahigh-frequency Techniques*, Van Nostrand, 1942.

Everitt, W. L., *Communication Engineering*, McGraw-Hill, 1937.

M.I.T. Staff, *Principles of Radar*, McGraw-Hill, 1946.

Sarbacher, R. I., and Edson, W. A., *Hyper- and Ultrahigh-frequency Engineering*, Wiley, 1943.

Skilling, H. H., *Fundamentals of Electric Waves*, McGraw-Hill, 1942.

Woodruff, I., *Electric Power Transmission*, Wiley, 1938.

CHAPTER 5

UHF APPLICATIONS OF LINES

If the only use for transmission lines in UHF communication circuits arose in the conveyance of energy from one point to another, there would be little justification for the emphasis given them in a text of this character. Fortunately, the characteristics of resonant lengths of a transmission line lend themselves to a comparatively wide field of application, including functions such as impedance matching, impedance transformation, and as circuit elements.

In the bulk of the applications at ultrahigh frequencies, the attenuation introduced is practically negligible and can, in many cases, be neglected. Certain conditions arise, particularly in the lower RF part of the electromagnetic spectrum, under which this approximation is not valid, necessitating the exercise of extreme vigilance in order to avoid improper use of relationships and concepts developed in connection with lossless lines.

Consideration of the dissipationless line leads to the development of generalized impedance and admittance diagrams, commonly called circle diagrams, which considerably enhance the solution of transmission line problems. This is frequently offset by the disadvantage arising from the fact that use of these charts can be made without proper understanding of their origin or significance, resulting in the learning of details of procedure rather than understanding of phenomena. It is a natural tendency which must be diligently discouraged.

5.1 Functions of Lines

One normally thinks of conveyance of electrical energy when considering the function of transmission lines, and certainly this is a very important application which they must fulfill. In radio communication systems of all types, some means of transferring the energy from transmitter to antenna, and from antenna to receiver is required. In radar, an added requirement is introduced since the same antenna is used for transmission and reception in most cases.

Consequently, a rather elaborate switching arrangement is necessitated for proper system operation.

The RF lines necessary to guide the electrical energy to the proper distribution points may be physically long or short, depending upon the relative locations of the transmitter or receiver and the antenna. However, due to the short wavelengths involved, these lines are electrically long, that is, they are an appreciable fraction, or even several multiples of a wavelength.

Fortunately, transmission lines have other very important properties. It will be shown in Art. 5.11 that a quarter wavelength of transmission line, short-circuited at one end, behaves like a parallel-tuned circuit at resonance. Above resonance it appears capacitive, below resonance it appears inductive. Thus, in a great many applications, short lengths of transmission line can be used as circuit elements, such as tuned circuits in oscillators. This property is highly advantageous since, at ultrahigh frequencies, it becomes physically impracticable to construct lumped inductances and capacitances which can be made to resonate at the radio frequencies normally encountered in UHF systems.

Associated with this tuned-circuit property of resonant lengths of transmission line is its impedance transforming characteristic which is extremely useful in impedance matching. Such lengths of line, normally a quarter or an eighth wavelength long, are called *transformers.*

Short lengths of low loss, short-circuited lines appear to have an input impedance which is almost purely reactive. Consequently,

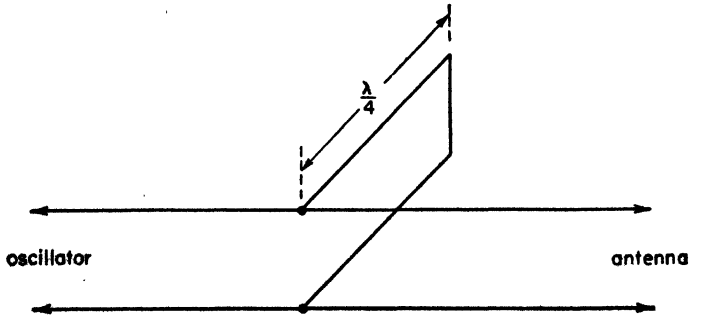


Fig. 5.1. Suppression of even harmonics.

such lengths of line are used to neutralize reactances along transmission line systems in order to match impedances and reduce the standing-wave ratio. These short lengths of line are called *stubs* and will be treated in detail in Arts. 5.13 and 5.14.

5.2 Supression of Even and Third Harmonics

The presence of harmonics of the oscillator frequency is undesirable (and illegal in many cases) and furnishes a situation to which the properties of resonant transmission lines may be applied to effect a remedy. For example, assume that even harmonics in the output of an oscillator are to be suppressed. This is readily accomplished by the arrangement shown in Fig. 5.1. A section of transmission line, called a "stub," shorted at one end, essentially lossless, and a quarter wavelength long at the fundamental frequency, is connected directly across the main transmission line joining the oscillator with the antenna. At the fundamental frequency, the input impedance of the stub is ideally infinite (practically, it is of the order of several hundred thousand ohms) so that it has no appreciable effect upon the fundamental. However, the wavelength of the second harmonic is

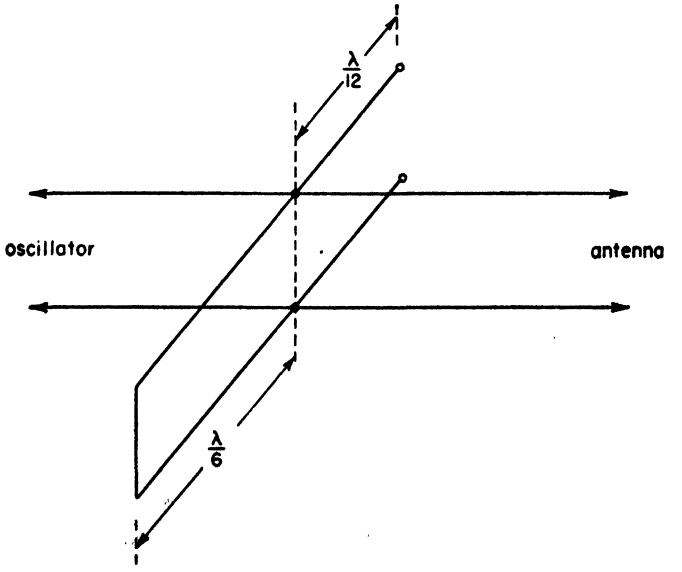


Fig. 5.2. Suppression of third harmonics.

half that of the fundamental, so that the stub is a half wavelength long at the second harmonic frequency. Moreover, it is always a multiple of a half wavelength at any even harmonic frequency. The input impedance of the stub is then zero, at these frequencies, and the unwanted second harmonics are shorted out before reaching the antenna.

Odd harmonics may be removed in a similar manner, but the connection is slightly different superficially. A typical connection for suppression of the third harmonic is shown in Fig. 5.2. A piece of transmission line, shorted at one end and open at the other, is connected as shown. It is a quarter wavelength long at the fundamental frequency, and the main transmission line is connected to it at a point a twelfth of a wavelength away from the open end. From the equation for the input impedance of a lossless line

$$Z_{\text{in}} = Z_c \frac{Z_R + jZ_c \tan \beta l}{Z_c + jZ_R \tan \beta l}$$

The input impedance of the open-circuited $\lambda/12$ section of the stub may be found as follows. Rearrange terms in the preceding equation, obtaining

$$Z_{\text{in}} = Z_c \frac{1 + j(Z_c/Z_R)\tan \beta l}{(Z_c/Z_R) + j \tan \beta l}$$

But, $Z_R = \infty$ = open circuit, so that

$$Z_{\text{in}} = \frac{Z_c}{j \tan \beta l} = -jZ_c \cot \beta l$$

Then letting $l = \lambda/12$, so that $\beta l = 30°$,

$$Z_{\text{in}} = -jZ_c \cot 30° = -j(1.732)Z_c$$

This is a capacitive reactance as indicated by the minus sign. The input impedance of the $\lambda/6$ shorted section is

$$Z_{\text{in}} = jZ_c \tan 60° = +j(1.732)Z_c$$

which is an inductive reactance exactly equal to the capacitive reactance of the open-circuited section. Consequently, the input impedance of the two sections in parallel is infinite at the fundamental frequency and the stub has no effect upon the signal. However, at the third harmonic, the open line is a quarter wavelength long and throws a short circuit directly across the main line. In addition, the

short-circuited section is a half wavelength long and also has an input impedance equal to zero. Thus, the third harmonic is effectively removed before reaching the antenna.

5.3 Measurement of Power

In general, the power input to a resistance may be expressed as

$$P = \frac{E^2}{R} = I^2R$$

On a lossless line, it was previously shown that the input impedance of the line at a voltage maximum is a pure resistance equal to ρZ_c, where ρ is the standing-wave ratio and Z_c is the characteristic impedance of the line. Hence, the power transmitted by the line may be written as

$$P = \frac{|E_{\max}|^2}{\rho Z_c} = |I_{\min}|^2 \rho Z_c$$

But, by definition, the standing-wave ratio is

$$\rho = \frac{|E_{\max}|}{|E_{\min}|} = \frac{|I_{\max}|}{|I_{\min}|}$$

Then substituting into the equation for the power,

$$P = \frac{|E_{\max}||E_{\min}|}{Z_c} = |I_{\max}||I_{\min}|Z_c$$

Thus, the power transmitted can be determined from an experimental determination of the voltage or current standing waves, if the characteristic impedance of the line is known. When the line is flat, that is, if there are no standing waves, then

$$P = \frac{|E|^2}{Z_c} = |I^2|Z_c$$

5.4 Measurement of Impedance

Assuming a section of lossless transmission line with a known characteristic impedance, then the value of any unknown impedance may be determined by connecting it as a load on the line and measuring the standing-wave pattern. For example, if the maximum value of the voltage standing wave appears across the load, then the unknown impedance is equal to ρZ_c, since at a voltage maximum,

$R_{in} = \rho Z_c$. If the minimum value of the voltage standing wave occurred at the load, then the unknown impedance would be Z_c/ρ.

For any intermediate condition, the calculation is made in the following manner. Let d = distance between the load and the first voltage maximum. At this point d, the input impedance is ρZ_c. Hence,

$$Z_{in} = \rho Z_c = Z_c \frac{Z_R + j_c Z \tan \beta d}{Z_c + jZ_R \tan \beta d}$$

Solve this equation for Z_R, the unknown impedance, obtaining

$$Z_R = Z_c\left(\frac{\rho - j \tan \beta d}{1 - j\rho \tan \beta d}\right)$$

If the measurement had been made to the first voltage minimum, then following the same procedure for the derivation yields

$$Z_R = Z_c\left(\frac{1 - j\rho \tan \beta d'}{\rho - j \tan \beta d'}\right)$$

where d' = distance from the load to the first voltage minimum.

5.5 Dissipationless Lines as Resonant Circuits

Resonant circuits to be used in the ultrahigh frequencies require very small values of inductance and capacitance. Generally speaking, it is impractical, if not impossible, to construct lumped-circuit parameters at these frequencies with such small values. In some cases it might be theoretically possible, but the elements would be so small that their power-handling capacity would be insufficient for most practical applications. Thus, a definite need exists for resonant circuits of some other type which will be physically larger so that they are easier to construct and tune as well as having larger power capacity.

Fortunately, certain properties of low-loss transmission lines may be utilized which have practically the same characteristics as resonant circuits. The general expression for the input impedance of a dissipationless transmission line is

$$Z = Z_c\left(\frac{Z_R + jZ_c \tan \beta l}{Z_c + jZ_R \tan \beta l}\right)$$

If the load impedance is a short circuit, then

$$Z_{sc} = jZ_c \tan \beta l$$

Whereas, if it is an open circuit,

$$Z_{oc} = -jZ_c \cot \beta l$$

The input impedance of such lines varies with the electrical length βl as shown in Fig. 5.3. The electrical length may be made to vary

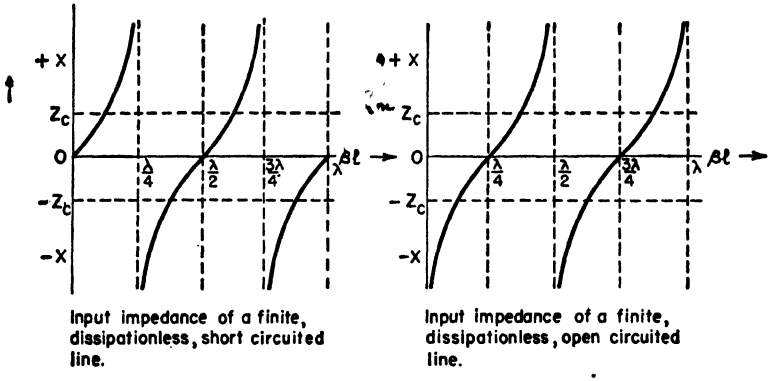


Fig. 5.3. Impedance characteristics of dissipationless lines.

by changing the frequency. Assuming the frequency to be the variable, it is evident that the impedance variations correspond rather closely to the series- and parallel-tuned circuits where

(1) Series resonant circuits correspond to:
  (a) odd multiples of a quarter wavelength of open line,
  (b) even multiples of a quarter wavelength of shorted line.
(2) Parallel resonant circuits correspond to:
  (a) even multiples of a quarter wavelength of open line,
  (b) odd multiples of a quarter wavelength of shorted line.

Resonant lengths of transmission line are sometimes used for interstage coupling in RF amplifiers. Furthermore, they are widely used as tuned circuits in high-frequency oscillators. However, calculations involved in circuit design are ordinarily more conveniently carried through in terms of ordinary lumped-circuit theory. Hence, it is highly desirable to establish an equivalence between the transmission line and the constants of a hypothetical tuned lumped-element circuit. The technique of determining the equivalent relationships is illustrated by the following example.

The input impedance of an open-circuited dissipationless line was shown to be

$$Z_{oc} = jX_{oc} = -jZ_c \cot \beta l$$

or

$$X_{oc} = -Z_c \cot \beta l$$

For a series-resonant circuit, the resonant frequency is defined by the relationship

$$X_0 = \omega_0 L - \frac{1}{\omega_0 C} = 0$$

Solving for the resonant frequency yields

$$\omega_0 = \sqrt{\frac{1}{LC}}$$

For the transmission line, if series resonance is to occur, then assuming an open-circuited line

$$X_{oc} = -Z_c \cot \beta_0 l = 0$$

This condition can be met only if

$$\cot \beta_0 l = 0$$

or, in terms of the angle, if

$$\beta_0 l = \frac{\pi}{2}$$

or any odd multiple thereof. Hence, solving for the phase constant β_0,

$$\beta_0 = \frac{\pi}{2l}$$

However, the phase constant was previously defined as

$$\beta_0 = \frac{2\pi}{\lambda_0} = \frac{2\pi f_0}{f_0 \lambda_0} = \frac{\omega_0}{v_0}$$

Thus, equating the two expressions for the phase constant yields

$$\frac{\omega_0}{v_0} = \frac{\pi}{2l}$$

Or, solving for the resonant frequency ω_0,

$$\omega_0 = \frac{\pi v_0}{2l}$$

Equating the expressions for resonant frequency yields

$$\frac{1}{\sqrt{LC}} = \frac{\pi v_0}{2l}$$

This gives one equation involving the two unknown constants of the hypothetical lumped-element circuit. A second equation is needed if both constants are to be evaluated.

If the tuned circuit is to closely correspond to the resonant line in the region of resonance, then the reactance curves should pass through zero with the same slope. These slopes are the derivatives of the reactances with respect to angular velocity, or $dX/d\omega$. For the transmission line

$$\frac{dX_{0c}}{d\omega} = \frac{d}{d\omega}(-Z_c \cot \beta l) = \frac{d}{d\omega}\left(-Z_c \cot \frac{\omega l}{v}\right)$$

Then carrying through the required derivative yields

$$\frac{dX_{0c}}{d\omega} = \frac{l}{v} Z_c \csc^2 \frac{\omega l}{v}$$

However, at resonance, it was shown that

$$\omega = \omega_0 = \frac{\pi v_0}{2l}$$

so that

$$\csc \frac{\omega_0 l}{v_0} = 1$$

Hence, the value of the slope of the reactance curve of the transmission line at the resonant frequency is

$$\left.\frac{dX_{0c}}{d\omega}\right|_{\omega=\omega_0} = \left(\frac{l}{v_0}\right)Z_c$$

Following the same procedure for the equivalent tuned circuit yields

$$\frac{dX}{d\omega} = \frac{d}{d\omega}\left(\omega L - \frac{1}{\omega C}\right) = L + \frac{1}{\omega^2 C}$$

At resonance the slope is

$$\left.\frac{dX}{d\omega}\right|_{\omega=\omega_0} = L + \frac{1}{\omega_0^2 C} = L + \frac{1}{\omega_0}\left(\frac{1}{\omega_0 C}\right) = L + \frac{\omega_0 L}{\omega_0} = 2L$$

Equating the values of the two slopes so obtained will produce

$$\left(\frac{l}{v_0}\right)Z_c = 2L$$

Solve for L and replace v_0 by $f_0\lambda_0$ to obtain

$$L = \frac{1}{2}\left(\frac{l}{\lambda_0}\right)\left(\frac{Z_c}{f_0}\right)$$

But, at resonance for the line, $l = \lambda/4$ so that

$$L = \frac{1}{2}\left(\frac{1}{4}\right)\frac{Z_c}{f_0} = \frac{1}{8}\left(\frac{Z_c}{f_0}\right) \tag{5.1}$$

This is the required value for L. The capacitance can be found from the equation for the resonant angular velocity

$$\omega_0 = \sqrt{\frac{1}{LC}}$$

Squaring both sides and substituting for L will give

$$\omega_0^2 = 4\pi^2 f_0^2 = \frac{1}{LC} = \frac{1}{\frac{1}{8}(Z_c/f_0)C} = \frac{8f_0}{Z_c C}$$

Then, solve for the capacitance to obtain

$$C = \left(\frac{2}{\pi^2}\right)\left(\frac{1}{Z_c f_0}\right) \tag{5.2}$$

The same procedure may be followed for the parallel equivalent, in which case the results are, for the short-circuited line,

$$C = \left(\frac{1}{8}\right)\left(\frac{1}{Z_c f_0}\right) \qquad L = \left(\frac{2}{\pi^2}\right)\left(\frac{Z_c}{f_0}\right) \tag{5.3}$$

One more circuit, mentioned in the article on harmonic suppression, deserves separate consideration. That is the quarter-wave line open at one end and shorted at the other. The total input impedance at any point is to be determined. Observe that it is always the sum of two impedances, in parallel; one capacitive, the open-circuited part, and one inductive, the short-circuited part. Consider the line shown in Fig. 5.4. If we break into the line at any point the input impedance is

$$Z = \frac{Z_{sc}Z_{oc}}{Z_{sc} + Z_{oc}}$$

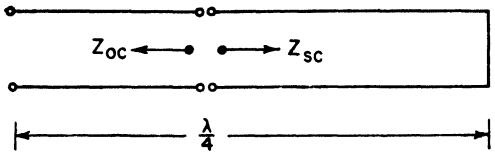


Fig. 5.4. Input impedance of a tapped quarter-wave line, open at one end, shorted at the other.

Then substituting for the impedances of the open- and short-circuited sections of the line yields

$$Z = \frac{jZ_c \tan \beta l\left[-jZ_c \cot\left(\frac{\lambda}{4} - l\right)\beta\right]}{jZ_c \tan \beta l - jZ_c \cot \beta\left(\frac{\lambda}{4} - l\right)}$$

By definition, the phase constant β is $2\pi/\lambda$. Thus, the cotangent term can be expanded by hyperbolic trigonometry, using the standard expansion, as follows:

$$\cot\left(\frac{2\pi}{\lambda}\right)\left(\frac{\lambda}{4} - l\right) = \cot\left(\frac{\pi}{2} - \frac{2\pi l}{\lambda}\right) = \tan \frac{2\pi}{\lambda} l = \tan \beta l$$

Consequently, the expression for the input impedance becomes

$$Z = \frac{(jZ_c \tan \beta l)(-jZ_c \tan \beta l)}{jZ_c \tan \beta l - jZ_c \tan \beta l} = \frac{Z_0^2 \tan^2 \beta l}{0} = \infty$$

Therefore, no matter where the break in the line is made, the input impedance is always ideally infinite. This occurs because the resonant frequency is determined by the product of L and C. Changing the position of the tap varies both L and C, but their product remains constant. The same effect is noted in the properties of lumped-element tapped resonant circuits.

5.6 Effective Q of Transmission Lines

If resonant lines are to be used as circuit elements, their Q is a factor which should be known. In general, the input impedance of a transmission line is given by the following equation:

$$Z = Z_c\left[\frac{Z_R + Z_c \tanh(\alpha + j\beta)l}{Z_c + Z_R \tanh(\alpha + j\beta)l}\right]$$

Assuming that the line is short-circuited at the load end, then

$$Z_{sc} = Z_c \tanh(\alpha + j\beta)l$$

Further, assume that the line is approximately a quarter wavelength long, being detuned from the $\lambda/4$ point by a small distance Δl. Then the total length is

$$l = l' + \Delta l = \frac{\lambda}{4} + \Delta l \qquad \text{so} \qquad \beta l' = \frac{\pi}{2}$$

Therefore, the input impedance in the short-circuited case becomes

$$Z_{sc} = Z_c \tanh\left(\alpha l + j\frac{\pi}{2} + j\,\Delta\beta l\right)$$

The hyperbolic tangent in this equation may be expanded into the form indicated in the following equation by use of a standard identity.

$$Z_{sc} = Z_c \frac{\tanh\,(\alpha + j\,\Delta\beta)l + j\tan\pi/2}{1 + [j\tanh\,(\alpha + j\,\Delta\beta)l][\tan\pi/2]}$$

Divide numerator and denominator through by $\tan\pi/2$ obtaining

$$Z_{sc} = Z_c \frac{j + \dfrac{\tanh\,(\alpha + j\,\Delta\beta)l}{\tan\pi/2}}{\left(\dfrac{1}{\tan\pi/2}\right) + j\tanh\,(\alpha + j\,\Delta\beta)l}$$

However, $\tan\pi/2 = \infty$, and consequently

$$Z_{sc} = \frac{Z_c}{\tanh\,(\alpha + j\,\Delta\beta)l}$$

Since the line was assumed to be physically short (about $\lambda/4$), then α, l, and Δl are all very small. As an approximation for small angles, the tangent and the angle are virtually equal so that

$$Z_{sc} \triangleq \frac{Z_c}{\alpha l + j\,\Delta\beta l}$$

The phase constant was previously defined as

$$\beta = \frac{2\pi}{\lambda} = \frac{2\pi f}{c} \qquad \text{where } c = \text{velocity of light}$$

Therefore, the input impedance reduces to

$$Z_{sc} \triangleq \frac{Z_c}{\alpha l + j(2\pi f/c)\Delta l}$$

Assuming that the transmission line is supplied by a constant-current generator, then the resonance curve will drop to the half-power points when the input impedance of the line is equal to $Z_0/\sqrt{2}$. This will occur only when the real and imaginary parts in the denominator of the equation for the input impedance are equal to one another. That is, when

$$\alpha l = \left(\frac{2\pi f}{c}\right)\Delta l = \left(\frac{2\pi l}{c}\right)\Delta f$$

Solving this equation for Δf yields

$$\Delta f = \frac{\alpha c}{2\pi}$$

The Q of the circuit may be defined from the frequency characteristic shown in Fig. 5.5.

$$Q = \frac{f_r}{2\,\Delta f}$$

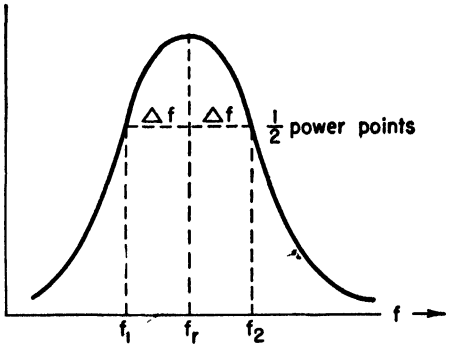


Fig. 5.5. Resonance curve.

Substituting for the Δf yields

$$Q = \frac{f_r}{2(\alpha c/2\pi)} = \left(\frac{\pi}{c}\right)\left(\frac{f_r}{\alpha}\right)$$

In the article on low-loss transmission lines it was shown that the expression for the attenuation constant had the approximate form

$$\alpha \doteq \frac{R}{2Z_c} + \frac{GZ_c}{2}$$

In general, however, the shunt conductance is so nearly zero that

$$\alpha \doteq \frac{R}{2Z_c}$$

Then, substituting for the attenuation gives

$$Q \doteq \left(\frac{2\pi f_r}{c}\right)\frac{Z_c}{R} = \left(\frac{\omega_r}{c}\right)\left(\frac{Z_c}{R}\right)$$

Calculation of the Q of a transmission line from this formula yields values ranging from 600 to as high as 30,000, assuming typical practical lines. The higher values of Q compare favorably with that found in crystals so that tuned lines may be used to stabilize oscillators for most UHF applications.

5.7 Derivation of the Rectangular Coordinate Circle Diagram

Through ordinary circuit analysis, electrical engineers are accustomed to the representation of impedances in a complex plane where resistance is plotted along the real, or X, axis, and the reactance is plotted along the imaginary, or Y, axis. The input impedance of a transmission line may be plotted in a similar manner, but with the added advantage that certain characteristics of the equation for the input impedance of a lossless line enable us to plot impedance (or admittance) loci in this complex plane, that are completely general. Such charts enormously facilitate the design of circuits using resonant lengths of line.

The input impedance of a transmission line is simply the ratio of the input voltage to the input current, that is

$$Z = \frac{V}{I} = \frac{E^+ + E^-}{I^+ + I^-}$$

Factor this equation as shown in the following expression.

$$Z = \left(\frac{E^+}{I^+}\right)\frac{1 + (E^-/E^+)}{1 + (I^-/I^+)}$$

But, the ratio of reflected to incident waves defines the reflection coefficient Γ where

$$\Gamma = \frac{E^-}{E^+} = -\frac{I^-}{I^+}$$

This may be rewritten in polar form as

$$|\Gamma| \underline{|-2\phi} = \frac{|E^-| \underline{|-\phi}}{|E^+| \underline{|+\phi}} = -\frac{|I^-| \underline{|-\phi}}{|I^+| \underline{|+\phi}}$$

Furthermore, the ratio of the incident-voltage wave to the incident-current wave equals the characteristic impedance of the line. Hence, the equation for the input impedance takes the following form:

$$Z = Z_c\left(\frac{1 + |\Gamma| \underline{|-2\phi}}{1 - |\Gamma| \underline{|-2\phi}}\right)$$

Divide through by the characteristic impedance Z_c, obtaining

$$z = \frac{Z}{Z_c} = \frac{1 + |\Gamma| \underline{|-2\phi}}{1 - |\Gamma| \underline{|-2\phi}}$$

This process is called *normalizing*, and all normalized impedances are customarily denoted by lower-case letters. That is

$$z = \frac{Z}{Z_c} = \frac{R}{Z_c} + j\frac{X}{Z_c} = r + jx$$

Now solve the equation for the normalized input impedance for the reflection coefficient.

$$|\Gamma| \underline{|-2\phi} = \frac{z - 1}{z + 1} = \frac{r + jx - 1}{r + jx + 1} = \frac{(r - 1) + jx}{(r + 1) + jx}$$

Rewriting the right-hand side of the equation in polar form yields

$$|\Gamma| \underline{|-2\phi} = \frac{\sqrt{(r-1)^2 + x^2}}{\sqrt{(r+1)^2 + x^2}} \underline{|\tan^{-1}\left(\frac{x}{r-1}\right) - \tan^{-1}\left(\frac{x}{r+1}\right)}$$

Equating magnitudes and angles on either side of the equality gives

$$|\Gamma| = \frac{\sqrt{(r-1)^2 + x^2}}{\sqrt{(r+1)^2 + x^2}}$$

$$-2\phi = \tan^{-1}\left(\frac{x}{r-1}\right) - \tan^{-1}\left(\frac{x}{r+1}\right) = \theta_1 - \theta_2$$

This process separates the two variables $|\Gamma|$ and 2ϕ, so that each can be analyzed separately.

Consider the equation for the magnitude of the reflection coefficient.

$$|\Gamma| = \frac{\sqrt{(r-1)^2 + x^2}}{\sqrt{(r+1)^2 + x^2}}$$

It was previously shown that the standing-wave ratio had the form

$$\rho = \frac{|\Gamma| + 1}{-|\Gamma| + 1}$$

Then, solving for the reflection coefficient

$$|\Gamma| = \frac{\rho - 1}{\rho + 1}$$

Therefore,

$$\frac{\rho - 1}{\rho + 1} = \sqrt{\frac{(r-1)^2 + x^2}{(r+1)^2 + x^2}}$$

Square both sides of this equation, rearrange in powers of r, and complete the square. This process yields

$$\left(r - \frac{\rho^2 + 1}{2\rho}\right)^2 + (x)^2 = \left(\frac{\rho^2 - 1}{2\rho}\right)^2$$

This is the standard form for the equation of a circle in the $r - x$ plane with a radius equal to $(\rho^2 - 1)/2\rho$ and with its center at

$$r = \frac{\rho^2 + 1}{2\rho} \qquad \text{and} \qquad x = 0$$

By inspection then, it is apparent that the center must lie completely on the real axis. Thus, typical plots appear as shown in Fig. 5.6. These circles were drawn by assuming some fixed value for the standing-wave ratio ρ. Consequently, they are called *constant-ρ circles*. These curves are especially interesting because the standing-

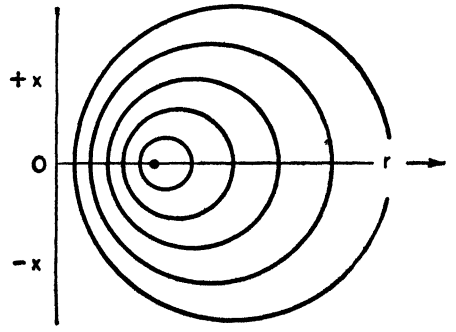


Fig. 5.6. Rectangular-coordinate circle diagram: constant-ρ circles.

wave ratio is a unique quantity, depending solely upon the comparative values of load and characteristic impedance. Thus, the standing-wave ratio characterizes any line with a given characteristic impedance and load.

These circles are plotted in what are called *chart units*, that is, all resistances and reactances were normalized by dividing them by the characteristic impedance of the line, which is a pure resistance on a lossless line. Thus, the coordinates of Fig. 5.6 are

$$r = \frac{R}{Z_c} \qquad \text{and} \qquad x = \frac{X}{Z_c}$$

It will be noted that the chart reactance is zero at the two points where the constant-ρ circles cross the chart resistance axis. At these points, $x = 0$, and therefore, from the equation for the circles

$$\left(r - \frac{\rho^2 + 1}{2\rho}\right) = \pm \frac{\rho^2 - 1}{2\rho}$$

Solving this equation for the two values of chart resistance yields

$$r_{\max} = \rho$$
$$r_{\min} = \frac{1}{\rho}$$

This is in accordance with the equations for the input impedance of a line at a voltage maximum or minimum. They were shown to be

$$R_{\max} = \rho Z_c \text{ or, in chart units, } r_{\max} = \rho$$
$$R_{\min} = \frac{Z_c}{\rho} \text{ or, in chart units, } r_{\min} = \frac{1}{\rho}$$

Hence, on the chart, the two points correspond to a voltage minimum and maximum. Since the maximum value of the chart resistance, for any given standing-wave ratio, equals ρ, then the value of ρ may be found directly from the chart for any load and line combination. If the input impedance of the line is given by

$$Z = R + jX$$

then normalize it to obtain

$$z = r + jx$$

and locate this point of the chart. It will fall on some constant-ρ circle. The value of ρ may be found immediately by following this

circle to the point where it crosses the chart resistance axis at its maximum value. The value of chart resistance read at that point is the magnitude of the standing-wave ratio.

To complete the chart, it is necessary to ascertain the nature of the lines of constant ϕ in the $r - x$ plane. It was previously shown that

$$-2\phi = \tan^{-1}\left(\frac{x}{r-1}\right) - \tan^{-1}\left(\frac{x}{r+1}\right) = \theta_1 - \theta_2$$

Now, take the tangent of both sides of this equation. That is,

$$\tan(-2\phi) = \tan(\theta_1 - \theta_2)$$

Or, using the standard trigonometric expansion on the right-hand side of the equation,

$$-\tan 2\phi = \frac{\tan\theta_1 - \tan\theta_2}{1 + \tan\theta_1 \tan\theta_2}$$

Substituting for θ_1 and θ_2 yields

$$-\tan 2\phi = \frac{\left(\frac{x}{r-1}\right) - \left(\frac{x}{r+1}\right)}{1 + \left(\frac{x}{r-1}\right)\left(\frac{x}{r+1}\right)} = \frac{2x}{r^2 + x^2 - 1}$$

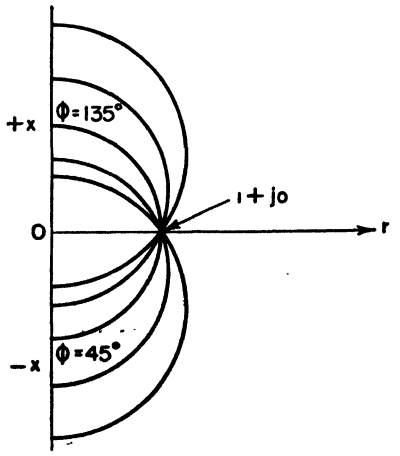


Fig. 5.7. Rectangular-coordinate circle diagram: constant-ϕ circles.

Rearranging terms and completing the square produces

$$r^2 + [x + \cot 2\phi]^2 = \frac{\tan^2 2\phi + 1}{\tan^2 2\phi} = \csc^2 2\phi$$

This is the equation of a circle in the $r - x$ plane specified in the following manner:

$$\text{Radius} = \csc 2\phi \qquad \text{center at } r = 0 \qquad \text{and} \qquad x = -\cot 2\phi$$

These circles appear as shown in Fig. 5.7. Since ϕ is constant along any one of these lines, they are appropriately called *constant-ϕ* circles. The complete chart then appears as shown in Fig. 5.10.

This is an impedance chart. However, precisely the same result would have been obtained for the input admittance of a line. In both cases, positive reactance and positive susceptance are in the top half of the chart. According to established convention,

$+jx$ = positive reactance = inductive reactance
$+jb$ = positive susceptance = capacitive susceptance

and conversely. Recognizing this fact, the chart may be used with equal facility, and in the same manner, for impedances or admittances.

5.8 Derivation of the Smith Chart

Although the rectangular coordinate circle diagram is a very valuable and easy-to-use tool, there is another chart which is generally better for ordinary use once the individual becomes accustomed to it. Because it is plotted in curvilinear coordinates, instead of rectangular coordinates, some difficulty may be experienced with it in early attempts at use. These troubles soon clear, however, and this *Smith chart** is usually preferred.

Commencing with the same expression used at the beginning of the derivation for the rectangular-coordinate circle diagram

$$Z = \frac{V}{I} = \frac{E^+ + E^-}{I^+ + I^-}$$

For the lossless line, and in exponential form, this becomes

$$Z = \frac{E^+ e^{j\beta l} + E^- e^{-j\beta l}}{I^+ e^{j\beta l} + I^- e^{-j\beta l}}$$

---

*P. H. Smith, Bell Telephone Laboratories, "A Transmission Line Calculator" page 29-31, *Electronics*, 12 (Jan. 1939).

From the definition of the reflection coefficient on a lossless line

$$\Gamma = \frac{E^-}{E^+} = -\frac{I^-}{I^+}$$

This may be substituted into the equation for the input impedance to obtain

$$Z = \left(\frac{E^+}{I^+}\right)\frac{e^{j\beta l} + \Gamma e^{-j\beta l}}{e^{j\beta l} - \Gamma e^{-j\beta l}}$$

The ratio of incident voltage to incident current is the characteristic impedance of the line. Consequently,

$$Z = Z_c\left(\frac{e^{j\beta l} + \Gamma e^{-j\beta l}}{e^{j\beta l} - \Gamma e^{-j\beta l}}\right)$$

Now multiply through numerator and denominator by $[e^{j\beta l} - \Gamma e^{-j\beta l}]$ to obtain

$$Z = Z_c\left(\frac{1 + \Gamma e^{-2j\beta l} - \Gamma e^{2j\beta l} - \Gamma^2}{1 - \Gamma e^{-2j\beta l} - \Gamma e^{2j\beta l} + \Gamma^2}\right)$$

$$= Z_c\left[\frac{(1 - \Gamma^2) - \Gamma(e^{2j\beta l} - e^{-2j\beta l})}{(1 + \Gamma^2) - \Gamma(e^{2j\beta l} + e^{-2j\beta l})}\right]$$

Using normalized impedances again, this may be written as

$$z = \frac{Z}{Z_c} = \frac{(1 - \Gamma^2) - 2j\Gamma \sin 2\beta l}{(1 + \Gamma^2) - 2\Gamma \cos 2\beta l} = r + jx$$

where $2j \sin 2\beta l = (e^{2j\beta l} - e^{-2j\beta l})$

and $2 \cos 2\beta l = (e^{2j\beta l} + e^{-2j\beta l})$

Then, equating real and imaginary terms on either side of the equation results in the following:

$$r = \frac{(1 - \Gamma^2)}{(1 + \Gamma^2) - 2\Gamma \cos 2\beta l}$$

$$x = \frac{2\Gamma \sin 2\beta l}{(1 + \Gamma^2) - 2\Gamma \cos 2\beta l}$$

Consider the expression for the normalized resistance r. Assume r to be fixed so that the reflection factor will have to vary as the electrical length βl is varied in order to maintain the equality. For any given value of normalized resistance assumed, a circle will be traced out, as shown in Fig. 5.8.

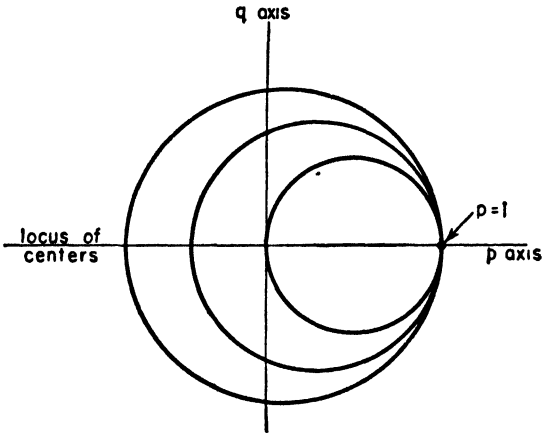


Fig. 5.8. Smith chart: constant-r circles.

To prove that the path traced out is a circle when the normalized resistance is held constant, rewrite the equation for r in rectangular coordinates by assuming that

$$p = \text{horizontal coordinate} = \Gamma \cos 2\beta l$$
$$q = \text{vertical coordinate} = \Gamma \sin 2\beta l$$

Hence,
$$p^2 + q^2 = \Gamma^2$$

Substitute these relationships into the equation for the normalized resistance. This process yields

$$r = \frac{1 - p^2 - q^2}{1 + p^2 + q^2 - 2p}$$

Cross multiplication reduces the equation to

$$r + rp^2 + rq^2 - 2rp = 1 - p^2 - q^2$$

Collect like powers of p and q, which will yield

$$p^2(1 + r) + q^2(1 + r) + p(-2r) = 1 - r$$

Divide through by $(1 + r)$,

$$p^2 + q^2 - p\left(\frac{2r}{1 + r}\right) = \left(\frac{1 - r}{1 + r}\right)$$

Complete the square in p to obtain

$$\left[p^2 - 2p\left(\frac{r}{1 + r}\right) + \left(\frac{r}{1 + r}\right)^2\right] + q^2 = \left(\frac{1 - r}{1 + r}\right) + \left(\frac{r}{1 + r}\right)^2$$

Hence,

$$\left(p - \frac{r}{1+r}\right)^2 + q^2 = \left(\frac{1}{1+r}\right)^2$$

This is the general form for the equation of a circle. The coordinates of the center are $p = r/(1 + r)$ and $q = 0$. The radius is $1/(1 + r)$. Note that the maximum radius is 1 when $r = 0$. When $q = 0$, the equation reduces to

$$\left(p - \frac{r}{1+r}\right)^2 = \left(\frac{1}{1+r}\right)^2$$

So that, taking plus signs throughout,

$$p - \frac{r}{1+r} = \frac{1}{1+r}$$

Hence, all circles of constant normalized r pass through the point 1 on the p axis. Therefore, although circles of different r have different radii, they all pass through the point ($p = 1, q = 0$) due to the change in location of the center of the circle. Thus, the constant-r circles appear as shown in Fig. 5.8.

A similar treatment of the normalized reactance component yields

$$(p - 1)^2 + \left(q \mp \frac{1}{x}\right)^2 = \left(\frac{1}{x}\right)^2$$

Hence, the lines of constant x are circles with centers at $p = 1$ and $q = 1/x$ while the radius is $1/x$. These constant-x circles are

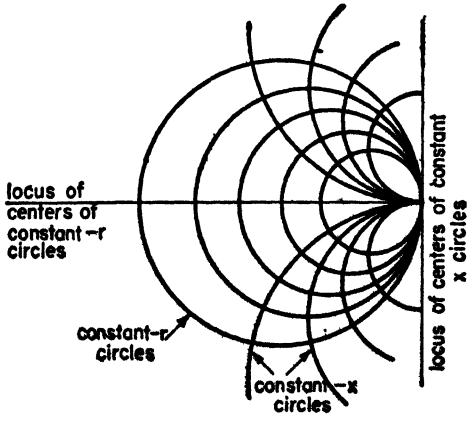


Fig. 5.9. Smith chart: theoretical form.

orthogonal to the constant-r circles at all points as shown in Fig. 5.9.

The useful part of the chart is confined to the region inside the constant-r circle for which $r = 0$. Thus, in practical form, the charts are circular in appearance.

If, in the equation for the normalized impedance, the reflection coefficient is held constant, while the normalized reactance and

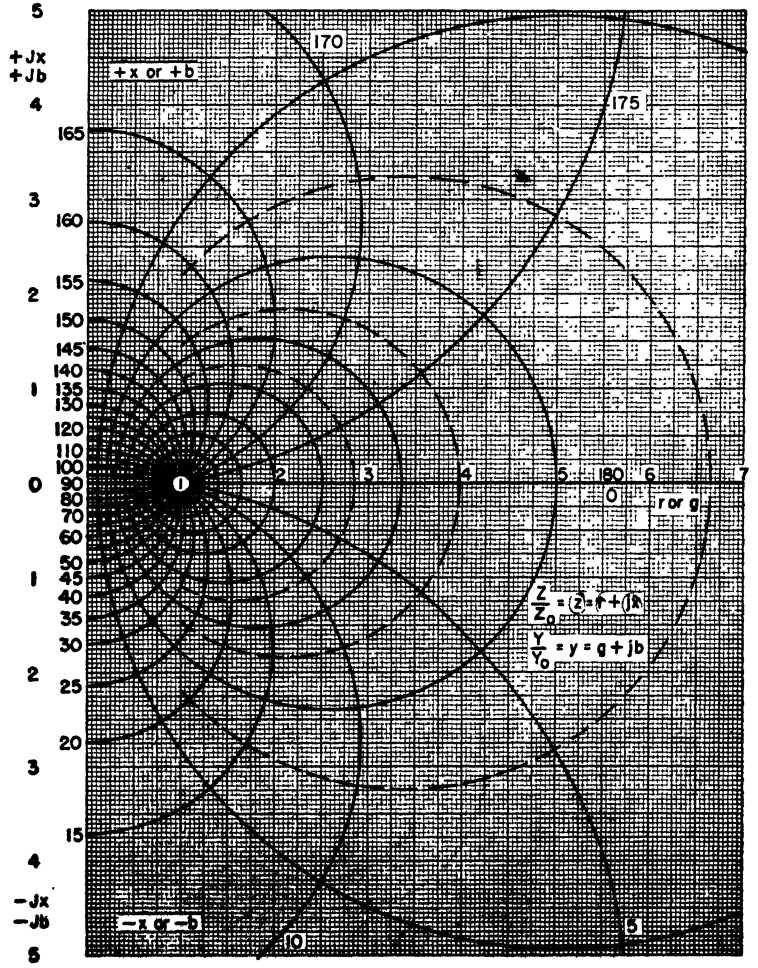


Fig. 5.10. Rectangular-coordinate circle diagram.

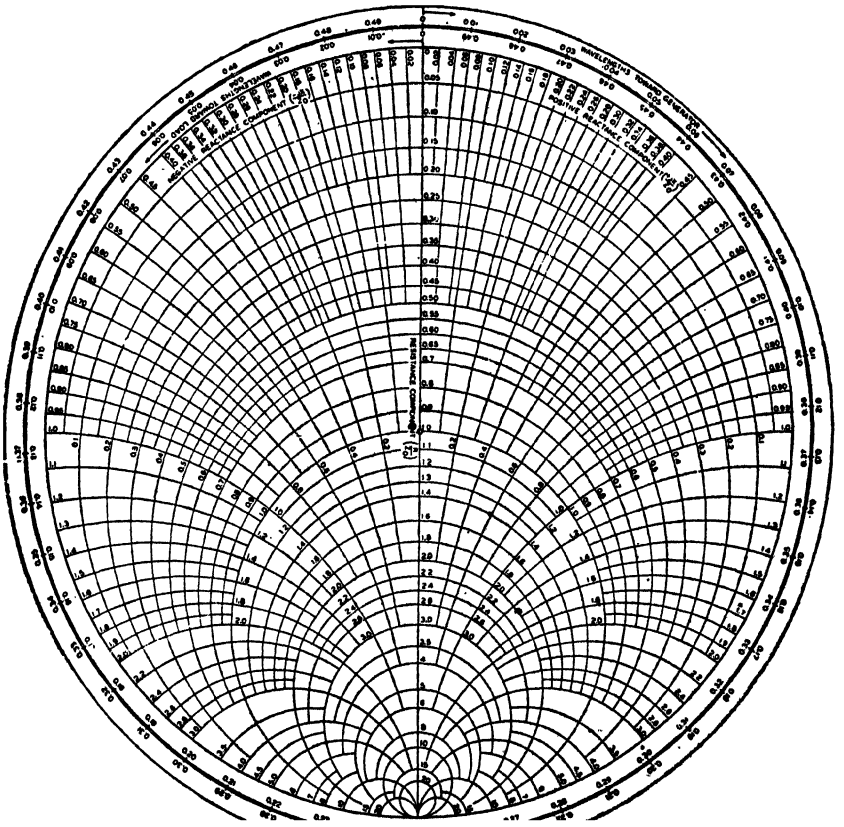


Fig. 5.11. Smith chart. (Courtesy P. H. Smith, Bell Telephone Laboratories.)

resistance are allowed to vary, a family of concentric circles with center at the origin of coordinates is obtained, referring to the p and q coordinate system. These are the constant-Γ or constant-ρ circles. Rotation through an electrical angle is made on these circles in the same manner as for the chart in rectangular coordinates. In passing, it should be noted that in many of the practical printed forms of the Smith chart, the constant-ρ circles are not printed on the figure, but must be drawn on for the case under consideration, using a compass.

This chart may be used with normalized admittances, too, just as in the case of the rectangular coordinate chart.

5.9 Comparison of the Two Diagrams

These charts, both the Smith chart and the circle diagram in rectangular coordinates, are purely general because of the normalizing process which removes the only term in the defining equations which specializes the charts to any one particular line. According to established convention, lower-case letters are used throughout to indicate normalized quantities, and the diagrams are then plotted in *chart units*, where 1 chart ohm = ohms/Z_c and 1 chart mho = mhos/Y_c.

The principal fact to remember in the application of these charts is that they are generalized vector diagrams, or more properly, generalized loci of impedance and admittance vector diagrams. Thus, the distinctions which are made between the two charts arise from the manner in which these loci are obtained, resulting in plots of the same quantities, but using different coordinate systems.

A visual comparison of the two charts shows that the following relationships may be established.

Locus of Constant	Circle Diagram in Rectangular Coordinates	Smith Chart
r or g	Vertical straight lines	Nonconcentric circles with centers on the horizontal diameter of the chart
x or b	Horizontal straight lines	Nonconcentric circles with centers on the vertical axis
ρ	Nonconcentric circles with centers on the chart resistance axis	Concentric circles about the center of the chart at the point $(1 + j0)$
ϕ	Nonconcentric circles with centers on the chart reactance axis	Straight lines from the chart center $(1 + j0)$ to the chart periphery

Since both charts yield the same information, merely using different methods of data presentation, discussion of one applies equally well to the other. Hence, for both charts

$r + jx$ denotes inductive impedance
$g + jb$ denotes capacitive admittance } top half of chart

$r - jx$ denotes capacitive impedance
$g - jb$ denotes inductive admittance } bottom half of chart

5.10 Input Impedance (or Admittance) of a Line

Despite frequent misconceptions to the contrary, these diagrams are used exclusively to determine the input impedance or admittance of a section of lossless line terminated in some load. Although they are used in the solution of comparatively difficult problems, such as double-stub matching, the actual function of the chart in every place where it enters the computation is merely to supply the value of the input impedance or admittance of a line. The charts are not oracles, or fabulous machines in which a flick of the wrist grinds out an answer. Their facility is restricted to the knowledge and common sense of the user. They simply make it unnecessary to go through the calculation of the input impedance equation

$$Z = Z_c\left(\frac{Z_R + jZ_c \tan \beta l}{Z_c + jZ_R \tan \beta l}\right)$$

at every step in a problem.

The only locus on these charts which is also a constant on any line and load combination is the standing-wave ratio. ρ is the same everywhere on the line. Likewise, ρ is the same everywhere on the constant-ρ circles. It is obvious then, that some direct relationship must exist between the point on the constant-ρ circle and the point, in terms of electrical length, on the line. It has already been shown that, on any one of the constant-ρ circles, the places where the circle crosses the normalized resistance axis correspond to the voltage maximum and minimum of the standing wave on the actual line. Since this electrical distance is a quarter of a wavelength, we conclude that there are 90 electrical degrees between these two points on the chart. Inspection of either chart shows this to be true. A half wavelength is then one complete revolution around the chart on a constant-ρ circle. Thus, any intermediate fraction of a wavelength can be scaled off by moving that $\Delta\phi$, as measured between the intersections of the appropriate constant-ϕ lines with the constant-ρ circles.

Following established convention, motion *away* from the load on the line corresponds to clockwise rotation (advancing phase) on the ρ circle. Motion toward the load corresponds to counter-clockwise rotation (retarding phase), on the chart ρ circle.

Assume that the input impedance of the line shown in Fig. 5.12 is to be determined. The load is

$$Z_R = R_R + jX_R$$

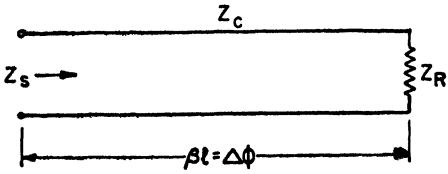


Fig. 5.12.

Normalize this impedance, obtaining

$$z_R = r_R + jx_R$$

Enter the chart with this value of impedance; that is, locate the point (r_R, x_R) on the chart. This point determines the standing-wave ratio characteristic of the line and load, as evidenced by the particular constant-ρ circle upon which the point falls. Furthermore, the point also lies on some constant-ϕ circle, ϕ_R. To find the input impedance Z_s, rotate clockwise (away from the load) on this constant-ρ circle, through an angle $\Delta\phi$ equal to the electrical length of the line in degrees. This locates the point ϕ_s where

$$\phi_s = \phi_R + \Delta\phi$$

The coordinates of this point are the coordinates of the input impedance in chart units. Read these values as

$$z_s = r_s + jx_s$$

The actual value of the input impedance is obtained by multiplying through by the characteristic impedance so that

$$Z_s = Z_c z_s = R_s + jx_s$$

The lines drawn from the center of coordinates of the chart to these points are the vectors z_R and z_s. Hence, the vector diagram obtained from the chart appears as shown in Fig. 5.13.

Exactly the same process is followed if the load admittance had been given and the input admittance was to be determined. Admittances are normalized in the same manner as impedances, that is, by dividing through by the characteristic admittance, or

$$y = \frac{Y}{Y_c} = \frac{G}{Y_c} + j\frac{B}{Y_c} = g + jb$$

If the input impedance had been given, and the load impedance was to be found, the procedure would be exactly reversed from that just outlined.

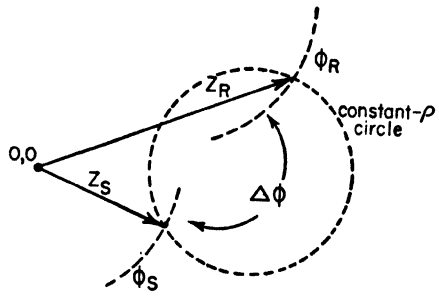


Fig. 5.13. Vector diagram from rectangular-coordinate circle diagram.

5.11 Impedance Transformation—Quarter-wave Section

The general expression for the input impedance of a lossless line is

$$Z_s = Z_c\left(\frac{Z_R + jZ_c \tan \beta l}{Z_c + jZ_R \tan \beta l}\right)$$

Divide numerator and denominator by tan βl

$$Z_s = Z_c\frac{(Z_R/\tan \beta l) + jZ_c}{(Z_c/\tan \beta l) + jZ_R}$$

For a quarter-wave section, $\beta l = 90°$, so tan $\beta l = \infty$, hence

$$Z_s = \frac{Z_c^2}{Z_R}$$

Because of this impedance-changing property indicated by the equation above, this is commonly called a *quarter-wave transformer*.

A typical application of this principle is found in matching an antenna to a transmission line. In this case, the antenna impedance is Z_R, and the input impedance to the quarter-wave section is designated by Z_s and must equal the characteristic impedance of the main transmission line. Since both Z_R and Z_s are known, the one remaining unknown is the characteristic impedance Z_c of the quarter-wave transformer. Solving the equation for Z_c yields

$$Z_c = \sqrt{Z_R Z_s}$$

The characteristic impedance of the quarter-wave section is adjusted to this calculated value by choosing the necessary wire spacing, diameter, and so on. A typical arrangement is shown in Fig. 5.14.

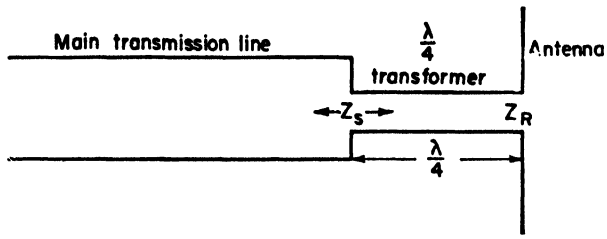


Fig. 5.14. Application of λ/4 transformer.

The equation showing the impedance transformation in a quarter-wave transformer is very interesting from the point of view of the circle diagrams. Since

$$Z_s = \frac{Z_c^2}{Z_R}$$

the terms may be rearranged as follows:

$$\frac{Z_s}{Z_c} = \frac{Z_c}{Z_R} = \frac{Y_R}{Y_c}$$

But

$$\frac{Z_s}{Z_c} = z_s = \text{normalized input impedance}$$

$$\frac{Y_R}{Y_c} = y_R = \text{normalized load admittance}$$

Thus, for a quarter-wave section, the normalized input impedance is equal to the normalized load admittance, That is, $z_s = y_R$. This gives us an extremely convenient method of converting impedances to admittances and vice versa, through the use of the circle diagrams. The procedure is simply to enter the chart at the given normalized impedance or admittance and rotate 90°. The coordinates of the resulting point give the value of the normalized admittance or impedance, respectively. For example, assume that the impedance $Z = R + jX$ is to be converted to admittance. The first step is to normalize Z to the form $z = r + jx$. Enter the chart at the point (r,x). From this point on the appropriate constant-ρ circle, rotate through an angle of 90°. Read the coordinates of this point as $y = g - jb$. Reversing the normalizing process yields

$$Y = Y_c y = G - jB$$

This is the desired result.

5.12 Impedance Transformation—Eighth-wave Section

The input impedance of a dissipationless line is

$$Z_s = Z_c\left(\frac{Z_R + jZ_c \tan \beta l}{Z_c + jZ_R \tan \beta l}\right)$$

If the line is an eighth of a wavelength long, $\beta l = 45°$ and $\tan \beta l = 1$.

So
$$Z_s = Z_c\frac{Z_R + jZ_c}{Z_c + jZ_R}$$

Or, in polar form,

$$|Z_s| \underline{\theta} = Z_c\frac{\sqrt{Z_R^2 + Z_c^2}\underline{|\tan^{-1} Z_c/Z_R}}{\sqrt{Z_c^2 + Z_R^2}\underline{|\tan^{-1}Z_R/Z_c}} = Z_c \underline{|\tan^{-1}\frac{Z_c}{Z_R} - \tan^{-1}\frac{Z_R}{Z_c}}$$

It is possible to write these equations this way only if both Z_c and Z_R are pure resistances. Z_c is, of course, since the line was assumed to be dissipationless. However, it must also be assumed that the load is likewise resistive. Under these conditions, if we equate absolute magnitudes on either side of the equation, $|Z_s| = Z_c$. Thus, the magnitude of the input impedance of an eighth-wave section terminated in a pure resistance is always equal to the characteristic impedance of the line. Conversely, an eighth-wave section terminated in an impedance equal in magnitude to the characteristic impedance has a pure resistance input impedance.

These facts may be verified quickly from the circle diagrams. For example, consider an eighth-wave section terminated in any pure resistance R_R. Normalize this to obtain r_R and enter the chart at the point $(r_R, 0)$. Rotate clockwise (away from the load) on the appropriate constant-ρ circle through 45°, where

$$\rho = r_R \quad \text{if} \quad r_R > 1$$

or
$$\rho = \frac{1}{r_R} \quad \text{if} \quad r_R < 1$$

The coordinates of the resulting point are always on the 45° or 135° constant-ϕ circles, depending upon r_R. The length of the line, drawn from the origin of coordinates to this point, is the magnitude of the normalized input impedance. By inspection it is apparent that its length is 1 chart unit, which is Z_c ohms. The converse may be demonstrated in the same manner.

5.13 Impedance Matching—Single Stub

A problem of the greatest practical importance is the matching of some load, such as an antenna, to a transmission line in order to eliminate the standing waves as well as match impedances. It is usually desirable to do this in such a way that the matching section is readily adjustable and does not alter the general configuration of the main transmission system. On open-wire lines the most common method of accomplishing this is through the use of a stub, adjustable in position on the main line, and variable in length. A typical connection is shown in Fig. 5.15. These stubs are usually short-circuited, though there are notable exceptions.

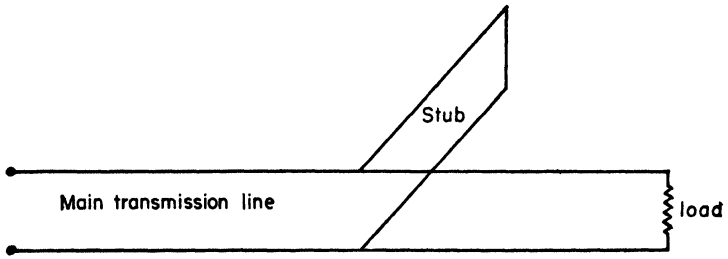


Fig. 5.15. Single stub.

Assuming that a given dissipationless line is terminated in a load impedance that is unequal to the characteristic impedance of the transmission line, then standing waves are produced due to reflections from the load. The input impedance of the line varies in a complex manner with the length, but at two points it is a pure resistance; a minimum (Z_c/ρ) at a voltage minimum, and a maximum (ρZ_c) at a voltage maximum. At some point on the line between these two extreme conditions, the resistive component of the input impedance equals the characteristic impedance, which is a pure resistance for the dissipationless line assumed. However, the input impedance has a reactive component which must be neutralized if an impedance match is to be made at this point. For such a match to occur, it is necessary to introduce this neutralizing reactance from some external source. The stub is that agency.

Because the stub is essentially lossless and terminated in an idealized "pure" open or short circuit, its input impedance must be a pure reactance. Since the stub cannot introduce any resistive component,

it must be *located* at a point on the line where the input impedance of the line has a resistance equal to the characteristic impedance of the line. The *length* of the stub is then adjusted so that it introduces the susceptance necessary to neutralize the susceptive component of the line input admittance. Since the stub is in parallel with the main line, computations are more conveniently carried through in terms of admittances than in impedances.

Although the design can be made directly from proper application of the equation for the input impedance or admittance of a line, the circle diagrams are decidedly more convenient. Consider the physical

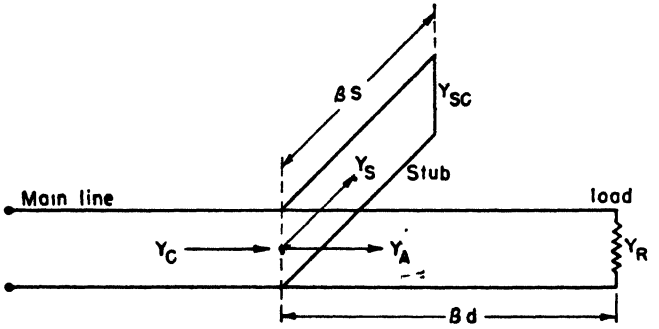


Fig. 5.16. Impedance matching with a single, movable stub.

arrangement shown in Fig. 5.16. For an impedance match to occur at the stub terminals,

$$Y_C = Y_A + Y_S$$

But, Y_C, the characteristic admittance, is a pure conductance G_C, while Y_S, the stub admittance, is a pure susceptance jB_S. Hence,

$$G_C = (G_A \pm jB_A) + (\pm jB_S)$$

Thus, the stub must be positioned so that

$$G_C = G_A$$

and have a length such that

$$B_S = -B_A$$

Vectorially, the problem appears as shown in Fig. 5.17.

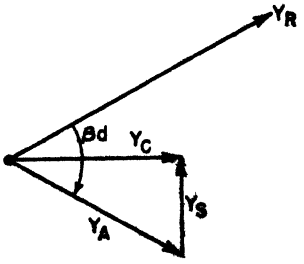


Fig. 5.17. Vector diagram for the single-stub matching setup.

Normalize all admittances. Note that

$$g_c = \frac{G_C}{Y_C} = 1 + j\,0 = 1$$

Hence,

$$y_A = 1 \pm jb_A$$

Draw the locus of all possible values of y_A which the stub can match to the line. On the circle diagrams this would simply be the line for which g is a constant and equal to 1 (g_c in normalized units). The susceptance can have any value. As long as y_A lies on this line, a match can be obtained between load and line because:

(1) The conductive terms are already equal.
(2) The stub can introduce any required neutralizing susceptance from 0 to ∞, depending upon its length.

Thus, the problem is to find out what value βd must have so that the normalized input admittance y_A of that section of the line to the right of the stub terminals in Fig. 5.16, terminated in y_R, falls on this locus.

Enter the chart at y_R. Rotate clockwise on the appropriate ρ circle until it intersects the previously drawn locus of y_A. The coordinates of this point give

$$y_A = 1 \pm jb_A$$

and the angle turned off in the rotation around the constant-ρ circle is obviously βd, the electrical length between load and stub. Hence, the stub susceptance, in chart units, is

$$y_S = \mp jb_A$$

Now it is necessary to determine the stub length that will produce this required value of susceptance. Consider the stub as a separate transmission line terminated in a short circuit, so that its load admittance is infinity. Enter the chart with this value of admittance

$$y_{sc} = \infty$$

The constant-ρ circle for this admittance is

(1) The left hand vertical axis on the rectangular coordinate chart.
(2) The periphery of the Smith chart.

Rotate clockwise on this constant-ρ circle to the point $0 \mp jb_A$.

The angle turned out in this rotation is βs, the electrical length of the stub. Thus, the load is matched to the line and standing waves to the left of the stub terminals are eliminated.

Note that there are two possible solutions to the problem, one at each intersection of the constant-ρ circle for y_R with the locus of y_A. Ordinarily, unless other mechanical features necessary to the particular installation require it to be otherwise, the shortest stub length is customarily used in order to keep the losses at a minimum.

The same procedure would be followed in matching a generator to the characteristic admittance of a line, but caution should be exercised in determining the proper directions of rotation on the constant-ρ circles.

5.14 Impedance Matching—Double Stubs

Although the single movable stub is simple in application and performs effectively, conditions arise which render its use impracticable. A typical case is the coaxial line, because with such construction, it would be very difficult to use movable stubs. This difficulty is overcome by using two stubs, both fixed in position, but variable in length. The physical configuration for a coaxial line appears in Fig. 5.18. The problem is clarified somewhat by con-

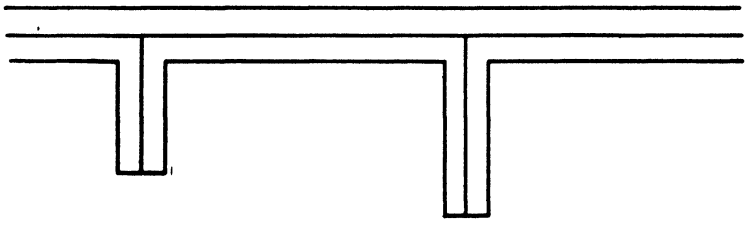


Fig. 5.18. Double-stub tuner on a coaxial line.

sidering the corresponding open-wire line, with double stubs, shown in Fig. 5.19. The distance between the two stubs is ordinarily fixed at either of three distances, an eighth, a quarter, or three-eighths of a wavelength. Inspection of Fig. 5.19 indicates that if a match is to occur between the line and the stub and load system, at the terminals of stub 2, then

$$Y_C = Y_B + Y_{S_2}$$

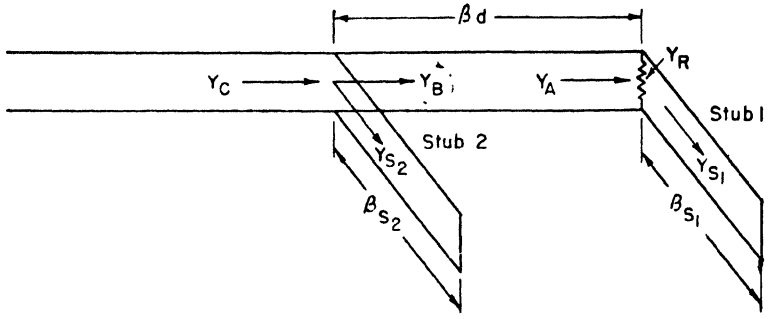


Fig. 5.19. Double-stub tuner on an open-wire line.

However, the characteristic admittance is a pure conductance, and the stub admittance is a pure susceptance, under the idealized conditions assumed. Therefore,

$$G_C = (G_B + jB_B) + jB_{S_2}$$

or, rearranging terms somewhat,

$$G_C = (G_B) + j(B_B + B_{S_2})$$

For an impedance match to be produced, corresponding real and imaginary terms on either side of the equation must be equal. That is,

$$G_B = G_C \qquad \text{and} \qquad B_B = -B_{S_2}$$

Consequently, the admittance looking into the line at the terminals of stub 2 is

$$Y_B = G_C \pm jB_B$$

and the stub admittance is

$$Y_{S_2} = \mp jB_B$$

Thus, the function of the second stub is to neutralize the susceptance of the line at that point. Therefore, Y_B, the input admittance of the line to the right of the terminals of stub 2, can have any susceptance, but its conductance *must* equal the characteristic admittance of the line. Thus, a restriction is placed upon Y_B. Moreover, Y_B is determined by the admittance Y_A and the electrical distance βd between stubs. But, since this distance is fixed, then Y_B is determined by Y_A. Consequently, Y_A is restricted in the sense that its value must

cause the conductive component of Y_B to equal the characteristic conductance. However

$$Y_A = Y_R + Y_{S_1}$$

Y_R, the load, is ordinarily fixed, so that only the *susceptance* of Y_A may be varied, by changing the length of stub 1. Thus, stub 1 functions in such a way that it causes Y_A to have a value which makes the conductance of Y_B equal to the characteristic admittance of the line. Stub 2 causes the net susceptance of Y_B to be neutralized, thus effecting the required impedance match. Thus, motion of a single stub is replaced here by the adjustment of a fixed stub (stub 1); thereafter the problem is the same as for a single stub. The argument may be more apparent from the vector diagram of Fig. 5.20.

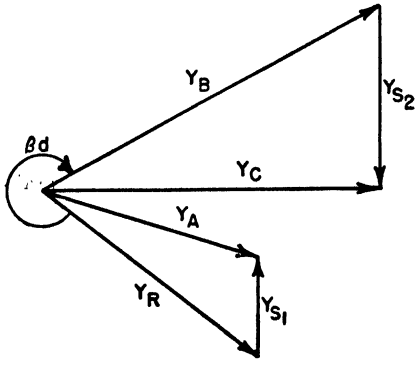


Fig. 5.20. Vector diagram of double-stub tuner operations.

The procedure for using the chart is as follows. Normalize all admittances. Draw the locus of y_B on the chart. This locus represents all the possible values of admittance which can be made equal to the characteristic admittance by adjustment of the length of stub 2. Since the stub can introduce susceptance only, the locus is obviously the line $g = 1$, just as in the case of the single stub.

Now it is necessary to find out what values of y_A will produce a y_B which falls on the locus just drawn. In other words, it is required to construct the locus of all values of y_A which produce values of y_B which adjustment of stub 2 can match to the line. This locus is constructed by arbitrarily assuming a series of points on the locus of y_B, already drawn, and rotating counterclockwise (toward the load) on the appropriate constant-ρ circle through the electrical distance βd between stubs. The resulting curve connecting all of the end points so obtained will be a circle, which is the locus of y_A desired. For convenience, this will be referred to as the *matching circle*. The matching circle simply gives us all the possible values

of y_A for which a match can be obtained. These values may or may not actually be available by adjustment of stub 1.

Now draw the locus of the values of y_A which we have at our disposal by adjustment of stub 1. Since stub 1 can change the susceptance of y_A from 0 to $\pm\infty$, this locus is obviously the line g_R.

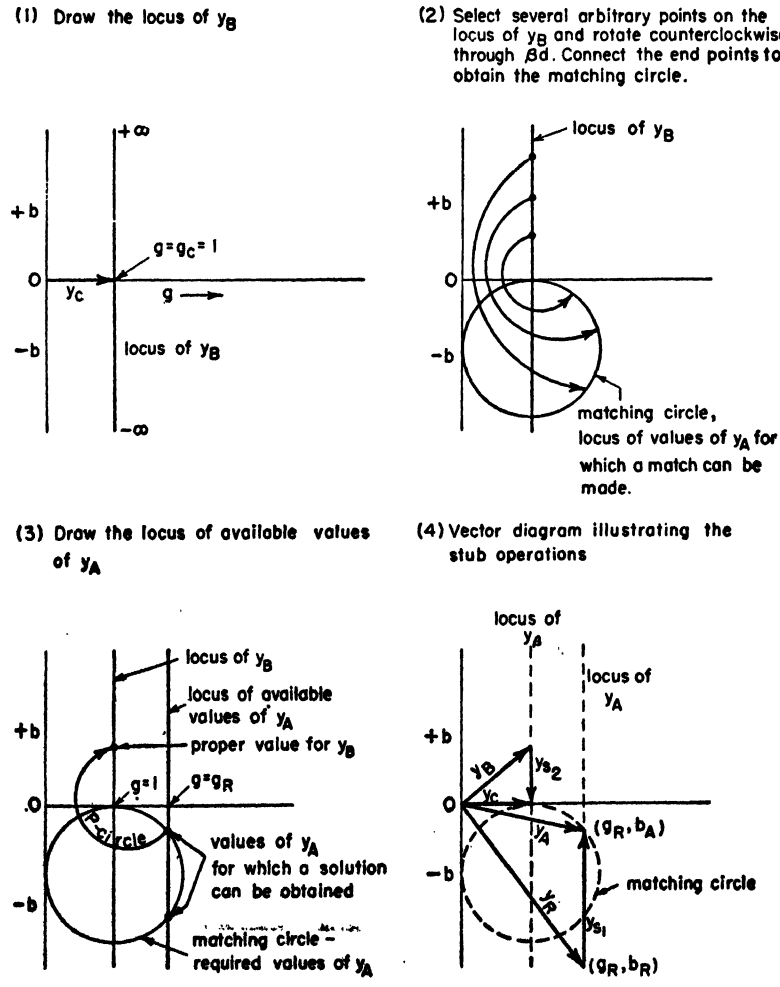


Fig. 5.21. Procedure for using the circle diagram in rectangular coordinates to design a double-stub matching system.

Where this locus of *available* values of y_A intersects the matching circle (the values of y_A for which a match can be obtained), is the value of y_A required by the problem. Since

$$y_A = g_R + jb_R + jb_{S_1} = g_A + jb_A$$

hence,

$$y_{S_1} = jb_{S_1} = j(b_A - b_R)$$

This gives the required value of the stub susceptance. Its length is found in the manner outlined in the preceding article for the single stub.

From the point y_A so determined, rotate clockwise on the appropriate constant-ρ circle through the angle βd. This locates a point on the previously drawn locus of y_B, thus yielding the value of y_B. The susceptance of stub 2 is then determined, since

$$y_{S_2} = jb_{S_2} = -jb_B$$

and the stub length is found in the usual manner. The steps in the solution of the double-stub problem are indicated, for the rectangular coordinate chart, in Fig. 5.21. It should be observed that, due to the method of construction, the matching circles are perfectly general for any given stub spacing, being entirely independent of the magnitude of the load admittance Y_R.

5.15 T-R And Anti T-R Devices

In most modern radar sets, transmitter and receiver are located at the same point and use the same antenna. The dual use of the antenna creates several special problems. The system would appear in its elemental form as shown in Fig. 5.22.

The transmitter generally places pulses of the order of 50 or 100 KW of peak power on the line. It is necessary to prevent this pulse from appearing at the receiver input terminals because it would cause serious damage to the crystal detector which is normally used. Even if detector damage were not a consideration, the possibility of receiver blocking is considerable. On the other hand, the echo pulse from the target is extremely weak and it is necessary to get as much of the

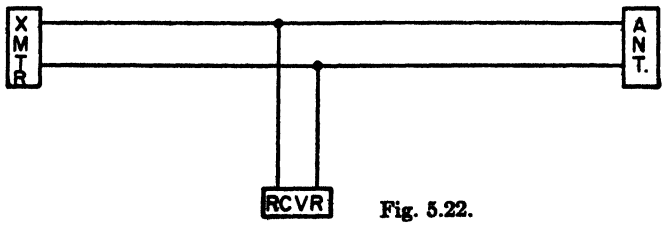


Fig. 5.22.

echo energy as possible to the receiver. Thus, during reception, there should be no absorption in the transmitter and the receiver should be matched to the line. In substance then, some device must be inserted in the RF system which

(1) disconnects the receiver from the line during transmission,
(2) disconnects the transmitter from the line during reception.

Such devices are called *Transmit-Receive (T-R) boxes.*

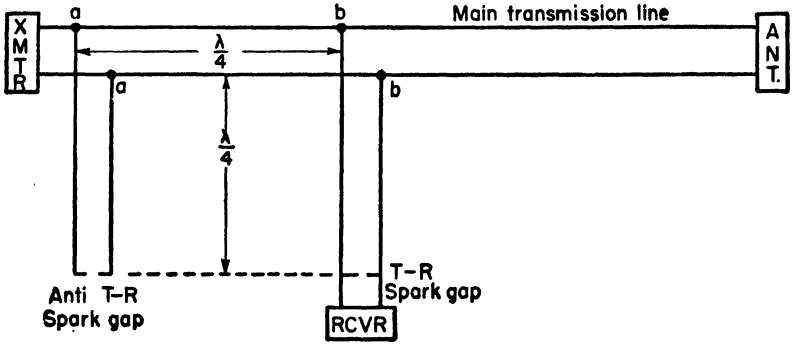


Fig. 5.23. Typical T-R, R-T system.

One method of accomplishing this is shown in Fig. 5.23. The high-voltage pulse from the transmitter causes both spark gaps to arc over. Ideally, these arcs are pure short circuits so that if it is assumed that the lines are dissipationless, the input impedance of both quarter-wave sections is infinite and both lines are disconnected from the main transmission line. Consequently, the pulse travels directly from transmitter to antenna.

During reception, the echo pulse is so weak that the spark gaps do not fire. Thus, the anti T-R line produces an effective short circuit at terminals *a-a*. At terminals *b-b* this appears as an open circuit so that the transmitter is effectively disconnected from the line. The T-R spark gap does not arc over and the receiver terminates the line in its characteristic impedance. Thus, all of the echo energy goes to the receiver.

Although this is only one of many methods, it illustrates the general technique required. The terminology T-R and Anti T-R is specified as follows:

(1) The T-R device disconnects and connects the receiver during transmission and reception, respectively.
(2) The Anti T-R, or R-T, device removes the effect of the transmitter impedance by disconnecting it during reception, but otherwise having no effect upon operation.

REFERENCES

Same as for Chapter 4.

PROBLEMS

5.1 Using the circle diagrams, find the input impedance of a lossless transmission line for which

$$Z_c = 100 \text{ ohms} \qquad Z_R = 200 + j150 \text{ ohms} \qquad Bs = 140°$$

Convert your result to admittance.

5.2 Find the input admittance of a lossless line if

$$Z_c = 50 \text{ ohms} \qquad Z_R = 50 - j100 \text{ ohms} \qquad Bs = 70°$$

5.3 The following data was obtained experimentally on a lossless transmission line:

$$\rho = 5 \qquad Z_c = 50 \text{ ohms} \qquad Bs = 70°$$

The distance from the load to the first voltage maximum is 20°. Find the load impedance. Find the input impedance.

5.4 Repeat Problem 5.3, assuming the distance from the load to the first voltage minimum is 20°.

5.5 A certain transmission line having a characteristic impedance of $100 + j0$ ohms is terminated in an admittance $Y_R = .04 - j.0195$ mhos. The line is 44 electrical degrees long. Find the length, in electrical degrees, of the short-circuited stub that must be added at the generator terminals to match the generator to the line if the internal admittance of the generator is $.005 + j0$ mhos.

5.6 A single, movable short-circuited stub is to be used to match a load Z_L to a transmission line. The following data apply:

$$Z_c = 50 + j0 \text{ ohms} \qquad Z_L = 12.5 + j25.0 \text{ ohms} \qquad \lambda = 3 \text{ cm}$$

Find the position (relative to the load) and length of the stub. Obtain both solutions.

5.7 A load admittance Y_L is to be matched to a transmission line by the double-stub method. The spacing between stubs is $3\lambda/8$. The first stub is located at the load terminals. If

$$Y_c = .01 + j0 \text{ mhos for line and stub}$$
$$Y_L = .0105 + j.0105 \text{ mhos}$$
$$f = 100 \text{ mcps}$$

Find the stub lengths. What range of conductances can the stub system match to the line?

5.8 An experimental transmission line, complete with double-stub tuner, is set up and terminated in an unknown impedance Z_L. By experiment it is found that the line is flat when the first stub is across the load terminals and 30 cm long. The second stub is 112.5 cm from the first stub and 45 cm long. If the characteristic impedance of line and stubs is $100 + j0$ ohms, and if the frequency is 100 mcps, find Z_L, assuming the stubs are short-circuited.

5.9 Repeat Problem 5.6, using a quarter-wave transformer. Find its two possible locations and the corresponding characteristic impedances.

5.10 An antenna is to be matched to a coaxial transmission line ($Z_c = 70$ ohms) by the double-stub method. The antenna is a folded dipole having an input impedance of $300 + j0$ ohms at the operating frequency of 100 mcps. The first stub is 37.5 cm back from the antenna and the stubs are $\lambda/4$ apart. Find the lengths of short-circuited stubs required. What range of conductances can the stub system match to the line?

5.11 Repeat Problem 5.10 for stubs $3\lambda/8$ apart.

5.12 Repeat Problem 5.10 for stubs $\lambda/4$ apart, but using open-circuited stubs.

5.13 What must be the characteristic impedance of a quarter-wave transformer to match a $(5000 + j0)$-ohm load to a generator having an internal resistance of 50 ohms? If the frequency is 300 mcps, and if the wire is No. 10 AWG (diameter = 0.1 inch), what should be the length and spacing of the wires for a two-wire matching section?

CHAPTER 6

PARALLEL PLANE WAVEGUIDES

It is a truism that professions become more sophisticated as their fields of application enlarge, and as the breadth and depth of the subjects with which they deal become more basic and profound. The physical sciences are certainly no exceptions. Every new device, formula, or technique passes through its period of novelty, surrounded by a "Buck Rogerish" atmosphere until it becomes integrated and accepted into the general field of knowledge to which it applies. Where the naive look for newness and differences, the sophisticate looks for similarities, and a profession becomes sophisticated when this process of integrating new devices and processes into the basic professional theory or philosophy is deliberately and actively pursued.

The phenomenon of wave guiding is not new to electrical engineers —a great many years have elapsed since electrical energy was conceived to flow like water inside the wires. Maxwell's field theory is conscientiously taught. Most engineers appreciate its significance and the fundamental physical and mathematical similarities of power and communication transmission lines, realizing that the distinction between the two arises from the practical applications which impose somewhat different requirements upon the design characteristics of the two systems.

Consequently, the engineer, who is an everlasting pragmatist, distinguishes only two fundamental classifications of electromagnetic waves: (1) guided and (2) unguided. *Guided* implies guiding during transmission, and such waves are characterized by those on any type of transmission line. The unguided wave is typified by a wave in free space, such as might have been radiated from an antenna. The directive properties of the antenna do not constitute guiding in the generic sense of the word. Under certain atmospheric conditions, and at certain frequencies, guiding of an otherwise free wave occurs between the ground and the atmosphere, but this action may be un-

avoidably produced. In any case, the guiding mechanism is not man-made. The introduction of guiding to an electromagnetic wave does not alter the fundamental nature of the phenomenon so that guided and unguided waves are closely related, although the surface distinctions appear very great indeed.

Waves may be guided in a great variety of ways, differing in detail, but depending upon the same fundamental effect; that is, any surface sharply separating two media of different dielectric constants exerts a guiding action on electromagnetic waves. This phenomenon in optics is a commonplace. Such discontinuities may be produced by either of two general methods:

(1) A surface between a conductor and an insulator.
(2) A surface between two insulators of distinctly different dielectric constants.

Thus, parallel-wire and coaxial transmission lines produce guiding by the first method, while dielectric rods and wires have been produced which utilize the second method.

It will be recalled that the coaxial line supplanted the open-wire line at high frequencies because of power loss considerations. In the material on transmission lines it was shown that the resistance of a single copper conductor is given by

$$R = 4.2\frac{\sqrt{\text{frequency}}}{\text{radius in centimeters}} \text{ microhms per meter}$$

This shows that, at high frequencies, the resistance is inversely proportional to radius of the wire. Thus, considering a parallel-wire system and a coaxial line of the same over-all dimensions, the copper loss in the coaxial line is much less since the inside radius of the outer conductor is much greater than the radius of each of the parallel wires. The loss in the outer conductor is very small, so that the loss in the inner conductor is of the greatest importance. If it were possible to transmit electromagnetic waves without the necessity of the inner conductor, the copper loss would be considerably reduced. The hollow pipe left when the inner conductor is removed is referred to, within the profession, as a *waveguide.*

That transmission in a hollow pipe is possible, under certain conditions, should not come as a profound shock to anyone nominally conversant with electromagnetic phenomena. The principle is the

same as that encountered in the normal course of study in transmission lines. Consider, for a moment, the rectangular hollow pipe shown in Fig. 6.1, which is constructed of some highly conducting material. Now assume that an electromagnetic wave is set up in the opening to the guide. Energy will tend to travel in all directions initially. However, energy traveling in the X and Y directions will be reflected by the short circuits formed by the walls of the pipe, so that, in the steady state, and if the X and Y dimensions are of the proper magnitude, standing waves of voltage or current will be set up across the face of the guide along the X and Y co-ordinates, and the only net transfer of energy will occur in the Z direction. The standing waves are necessary in order to meet the boundary conditions at the metal surfaces and it is necessary for the X and Y dimensions to be some critical fraction of a wavelength, depending upon the excitation, for this to occur. Thus, the dimensions of the waveguide are roughly of the magnitude of the wavelength, a fact which renders it impractical at frequencies below about 10,000 mcps.

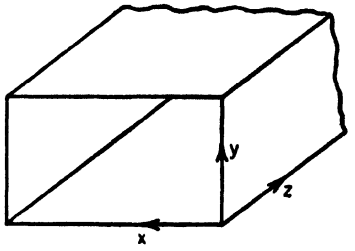


Fig. 6.1. Rectangular hollow pipe waveguide.

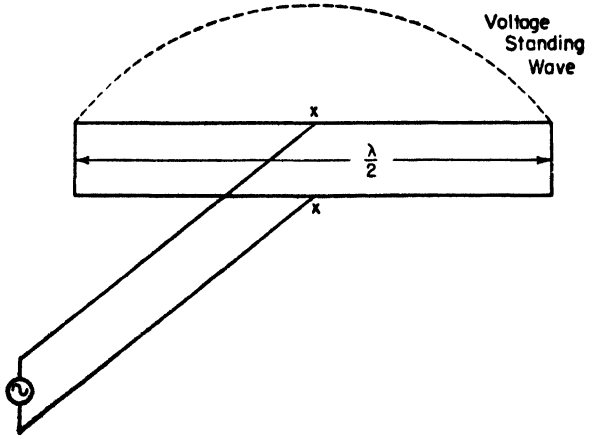


Fig. 6.2.

The relationship to transmission line study may be more apparent if the concept of waveguiding is approached in the following manner. Consider the setup shown in Fig. 6.2 where a generator is supplying a transmission line, shorted at both ends and a half wavelength long. Such a line appears as a parallel tuned circuit and a voltage standing wave appears as shown. If, at the terminals x-x, a length of transmission line terminated in its characteristic impedance is connected to the circuit, then, in the steady state, the wave travelling down this line is unaffected by the presence of any number of the short-circuited lines. Then considering the limiting case where an infinite number of lines of both types are used, we obtain the closed rectangular waveguide.

At the higher frequencies, waveguides have characteristics, other than low copper loss, which make them highly desirable. For example, at very high frequencies the dielectric loss in the bead supports of coaxial lines becomes objectionably large. Furthermore, the bead separation becomes an appreciable fraction of a wavelength and the resulting reflections, caused by these discontinuities, lower the efficiency of the system.

Still another consideration is the power-handling capacity of waveguides, which is much greater than for a coaxial line designed for use at the same frequency. The extreme ruggedness, ease of manufacture, and practical size at frequencies of 10,000 mcps and above are further points favoring the use of waveguides over coaxial lines.

By way of summary then, waveguides are generally used at frequencies of 10,000 mcps and above, since their cost at lower frequencies is prohibitive compared to coaxial lines. In the region around 3000 mcps, although waveguides were used at one time, coaxial line is almost always employed now.

6.1 Basic Considerations

The subject of waveguides may be approached in a wide variety of ways, but the simplest, and generally the most illuminating, is to effect a bridge between conventional transmission lines and practical waveguides that will relate the new to the familiar. The parallel plane waveguide, or parallel plane transmission line, affords an excellent "discussional bridge," to coin a phrase.

Consider two parallel plane conductors of infinite extent, separated by a dielectric, and oriented with respect to an assumed set of coordinate axes as shown in Fig. 6.3. Unless definitely specified otherwise, discussion in this chapter will be confined to dielectric mediums and metallic conductors which are homogeneous and isotropic, so that the dielectric constant, permeability, and conductivity (ϵ, μ, σ, respectively) are true constants. Under these assumptions, it will be recalled from fundamental electromagnetic theory that

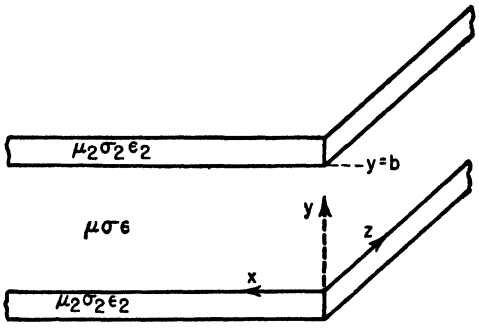


Fig. 6.3. Nomenclature associated with the parallel plane waveguide.

$$\mathbf{B} = \mu\mathbf{H} \tag{6.1}$$

$$\mathbf{D} = \epsilon\mathbf{E} \tag{6.2}$$

where $\mathbf{B}$ = magnetic induction, $\mathbf{D}$ = electric induction, $\mathbf{H}$ = magnetic field intensity, and $\mathbf{E}$ = electric field intensity. Maxwell's equations,

$$\text{Curl } \mathbf{H} = \sigma\mathbf{E} + \epsilon\frac{\partial \mathbf{E}}{\partial t} \tag{6.3}$$

$$\text{Curl } \mathbf{E} = -\mu\frac{\partial \mathbf{H}}{\partial t} \tag{6.4}$$

may be written in expanded form in rectangular (Cartesian) coordinates as shown in Eq. (6.5).

$$\left.\begin{aligned} \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} &= \sigma E_x + \epsilon\frac{\partial E_x}{\partial t} & \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} &= -\mu\frac{\partial H_x}{\partial t} \\ \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} &= \sigma E_y + \epsilon\frac{\partial E_y}{\partial t} & \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} &= -\mu\frac{\partial H_y}{\partial t} \\ \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= \sigma E_z + \epsilon\frac{\partial E_z}{\partial t} & \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} &= -\mu\frac{\partial H_z}{\partial t} \end{aligned}\right\} \tag{6.5}$$

Assume that under steady-state conditions, propagation is along the Z axis. Due to the boundary conditions imposed by the presence of the plane conductors normal to the Y axis, the magnetic and

electric field intensities are not constant with respect to distance measured along the Y coordinate. However, since the planes are of infinite extent, no such boundary conditions need be imposed along the X axis, so that no variation in either of the intensity vectors will occur in this direction. In other words

$$\frac{\partial H_z}{\partial x} = \frac{\partial H_y}{\partial x} = 0 \qquad \frac{\partial E_z}{\partial x} = \frac{\partial E_y}{\partial x} = 0$$

Consequently, Maxwell's equations, given by Eq. (6.5) reduce to

$$\left.\begin{aligned} \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} &= \sigma E_x + \epsilon\frac{\partial E_x}{\partial t} & \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} &= -\mu\frac{\partial H_x}{\partial t} \\ \frac{\partial H_x}{\partial z} &= \sigma E_y + \epsilon\frac{\partial E_y}{\partial t} & \frac{\partial E_x}{\partial z} &= -\mu\frac{\partial H_y}{\partial t} \\ -\frac{\partial H_x}{\partial y} &= \sigma E_z + \epsilon\frac{\partial E_z}{\partial t} & -\frac{\partial E_x}{\partial y} &= -\mu\frac{\partial H_z}{\partial t} \end{aligned}\right\} \quad (6.6)$$

Most electromagnetic signals with which one customarily deals are periodic functions of time. Furthermore, by means of the Fourier series method of analysis, any periodic function may be resolved into a series composed of sine and cosine terms, each such term being made up of factors of the general form $e^{j\omega t}$. One particular solution of the differential equation will be obtained for each such term, and since the differential equations are linear and homogeneous, the principle of superposition can be used to obtain the complete solution by summing up all of the individual particular solutions. Hence, the assumption that the field intensities are harmonic functions of time, as given by the term $e^{j\omega t}$, does not reduce the generality of the treatment materially. So, assume that the instantaneous values of $\mathbf{E}$ and $\mathbf{H}$ are given by

$$\left.\begin{aligned} \mathbf{E} &= \mathbf{E}^m e^{j\omega t} \quad \text{where } \mathbf{E}^m = \text{complex magnitude of } \mathbf{E} \\ \mathbf{H} &= \mathbf{H}^m e^{j\omega t} \quad \text{where } \mathbf{H}^m = \text{complex magnitude of } \mathbf{H} \end{aligned}\right\} \quad (6.7)$$

and substitute these expressions into Maxwell's equations for the system under consideration, as given by Eq. (6.6). This process yields

$$\left.\begin{aligned} \frac{\partial H_z^m}{\partial y} - \frac{\partial H_y^m}{\partial z} &= (\sigma + j\omega\epsilon)E_x^m & \frac{\partial E_z^m}{\partial y} - \frac{\partial E_y^m}{\partial z} &= -j\omega\mu H_x^m \\ \frac{\partial H_x^m}{\partial z} &= (\sigma + j\omega\epsilon)E_y^m & \frac{\partial E_x^m}{\partial z} &= -j\omega\mu H_y^m \\ -\frac{\partial H_x^m}{\partial y} &= (\sigma + j\omega\epsilon)E_z^m & -\frac{\partial E_x^m}{\partial y} &= -j\omega\mu H_z^m \end{aligned}\right\} \quad (6.8)$$

6.2 Modes of Transmission

Equations (6.8) may be reduced into two fundamentally different sets of equations, depending upon the manner in which the generator orients the field in the waveguide. The two distinct cases arise since it is possible to produce a field in the guide such that

(1) the electric field intensity vector has no Y or Z component' that is, it is along the X axis and has only an X component, or
(2) the magnetic field intensity vector is along the X axis and has no Y or Z components.

Two fundamentally different field configurations result.

Consider the first case. Assume that the electric field intensity **E** is parallel to the X axis. Under this assumption **E** has only one component E_x; the others are zero. Hence, Eqs. (6.8) reduce to

$$\left.\begin{aligned} &\text{(a)}\ \frac{\partial H_z^m}{\partial y} - \frac{\partial H_y^m}{\partial z} = (\sigma + j\omega\epsilon)E_x^m \qquad \text{(d)}\ 0 = -j\omega\mu H_x^m \\ &\text{(b)}\ \frac{\partial H_x^m}{\partial z} = 0 \qquad\qquad\qquad\qquad \text{(e)}\ \frac{\partial E_x^m}{\partial z} = -j\omega\mu H_y^m \\ &\text{(c)}\ \frac{\partial H_x^m}{\partial y} = 0 \qquad\qquad\qquad\qquad \text{(f)}\ -\frac{\partial E_x^m}{\partial y} = -j\omega\mu H_z^m \end{aligned}\right\} \quad (6.9)$$

According to Eq. (6.9-d)

$$-j\omega\mu H_x^m = 0$$

Except in the trivial case, the $(j\omega\mu)$ cannot be zero so that we conclude that

$$H_x^m = 0$$

But, the other two components of the magnetic field intensity, H_y^m and H_z^m, are not zero. Then the only components of the field intensities involved in this mode of transmission appear as shown in Fig. 6.4.

Since H_x^m is zero, the magnetic field intensity vector lies entirely within the Y-Z plane, and tends to be tipped along Z in the direction of propagation. On the other hand, the electric field intensity vector lies directly along the X axis and is accordingly entirely transverse to the direction of propagation. Consequently, this field configuration is commonly referred to as the *transverse electric* (*TE*) *mode* of transmission.

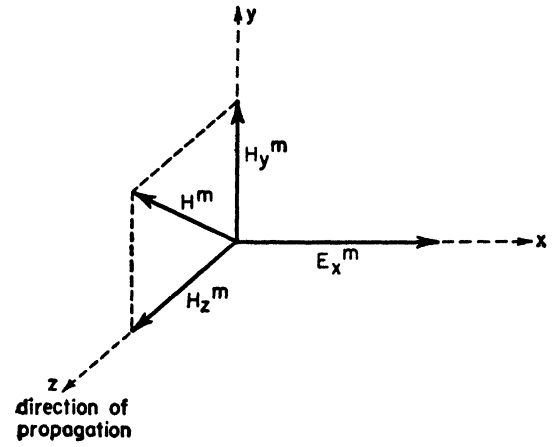


Fig. 6.4. Field intensity components in the TE mode of transmission.

On the other hand, consider the second case in which the magnetic field was assumed to be entirely parallel to the X axis. Then it would have only one component H_x and Eqs. (6.8) would reduce to

$$\left.\begin{array}{ll} \text{(a)}\ 0 = (\sigma + j\omega\epsilon)E_x^m & \text{(d)}\ \dfrac{\partial E_z^m}{\partial y} - \dfrac{\partial E_y^m}{\partial z} = -j\omega\mu H_x^m \\ \text{(b)}\ \dfrac{\partial H_x^m}{\partial z} = (\sigma + j\omega\epsilon)E_y^m & \text{(e)}\ \dfrac{\partial E_x^m}{\partial z} = 0 \\ \text{(c)}\ -\dfrac{\partial H_x^m}{\partial y} = (\sigma + j\omega\epsilon)E_z^m & \text{(f)}\ -\dfrac{\partial E_x^m}{\partial y} = 0 \end{array}\right\} \quad (6.10)$$

According to Eq. (6.10-a)

$$(\sigma + j\omega\epsilon)E_x^m = 0$$

So $E_x^m = 0$ since $(\sigma + j\omega\epsilon)$ cannot be zero.
Hence, the electric field intensity vector has two components, E_y^m and E_z^m, which are not zero and the field configuration appears as shown in Fig. 6.5. Since the $\mathbf{H}^m$ vector is entirely along the X axis, it is transverse to the direction of propagation, Z, and the field configuration is termed the *transverse magnetic* (*TM*) *mode.*

There is another way of designating these modes, common to many texts, but apparently slowly going out of usage, which specifies the mode according to the field having a component in the direction of propagation, that is, along the Z axis. For example, in the TE

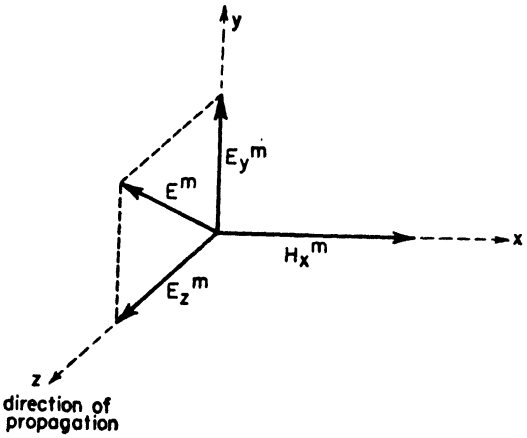


Fig. 6. 5. Field intensity components in the TM mode of transmission.

mode, the **H** vector has a component in the Z direction, whereas the **E** vector does not. Consequently, it is called an H mode. Conversely, and for the same reason, the TM mode is sometimes termed an E mode. In each case the letter refers to an "electric" type or "magnetic type" mode. This designation, however, will *not* be used in this book and is mentioned in passing only to acquaint the reader with the terminology that may sometimes be encountered.

6.3 Derivation of the Wave Equation

The solutions of the differential equations obtained in the preceding articles are most directly found by means of the *wave equations*, equations which appear in all branches of physics, in various forms differing in detail only, where wave motion occurs. The wave equations are obtained directly from Maxwell's equations by separating variables, and obtaining expressions containing **E** and **H** separately instead of together in the same equation. Thus, the wave equations are merely restatements of Maxwell's equations in a more mathematically convenient form.

In vector notation, Maxwell's equations may be written as

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \tag{6.11}$$

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t} + \sigma \mathbf{E} \tag{6.12}$$

Now take the curl of Eq. (6.11), obtaining

$$\nabla \times (\nabla \times \mathbf{E}) = -\nabla \times \left(\mu \frac{\partial \mathbf{H}}{\partial t}\right) \tag{6.13}$$

However, by a vector identity, the left-hand side of Eq. (6.13) may be written as

$$\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} \tag{6.14}$$

The quantity $\nabla \cdot \mathbf{E}$ is the divergence of the electric field intensity. In Art. 6.1 it was assumed that the dielectric was homogeneous. Consequently, the divergence of either $\mathbf{E}$ or $\mathbf{H}$ must be zero. Hence, Eq. (6.14) reduces to

$$\nabla \times \nabla \times \mathbf{E} = -\nabla^2 \mathbf{E} \tag{6.15}$$

Substitution of Eq. (6.15) back into (6.13) yields

$$\nabla^2 \mathbf{E} = \nabla \times \left(\mu \frac{\partial \mathbf{H}}{\partial t}\right) \tag{6.16}$$

Since the permeability μ has been assumed to be a constant it can be brought outside the brackets, giving

$$\nabla^2 \mathbf{E} = \mu \frac{\partial}{\partial t} (\nabla \times \mathbf{H}) \tag{6.17}$$

The quantity now remaining in the brackets is the curl $\mathbf{H}$, which, by Eq. (6.12) would make Eq. (6.17) become

$$\nabla^2 \mathbf{E} = \mu \frac{\partial}{\partial t}\left(\sigma \mathbf{E} + \epsilon \frac{\partial \mathbf{E}}{\partial t}\right) \tag{6.18}$$

Carrying out the indicated differentiations yields

$$\nabla^2 \mathbf{E} - \mu\sigma \frac{\partial \mathbf{E}}{\partial t} - \mu\epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0 \tag{6.19}$$

In operational form this may be written as

$$\left(\nabla^2 - \mu\sigma \frac{\partial}{\partial t} - \mu\epsilon \frac{\partial^2}{\partial t^2}\right)\mathbf{E} = 0 \tag{6.20}$$

Following the same procedure for the curl $\mathbf{H}$ would produce

$$\left(\nabla^2 - \mu\sigma \frac{\partial}{\partial t} - \mu\epsilon \frac{\partial^2}{\partial t^2}\right)\mathbf{H} = 0 \tag{6.21}$$

In both cases, the quantity ∇^2 is the *Laplacian*, which is defined as

$$\nabla^2 \triangleq \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{6.22}$$

Equations (6.20) and (6.21) are the wave equations. They must be satisfied by each component of the electric or magnetic field intensities.

It is interesting to note that when the conductivity of the dielectric is zero, the wave equations reduce to the form

$$\left(\nabla^2 - \mu\epsilon \frac{\partial^2}{\partial t^2}\right)\mathbf{E} = 0 \tag{6.23}$$

which is the wave equation for free space. The $\mu\sigma(\partial/\partial t)$ term appearing in the complete wave equation functions as a damping factor introducing attenuation.

In the particular case under consideration, it was assumed that both of the field intensities were harmonic functions of time. A simplification in the mathematics may be made by assuming that this time dependence is specified as follows:

$$\mathbf{E} = \mathbf{E}^m \sin \omega t \qquad \text{and} \qquad \mathbf{H} = \mathbf{H}^m \sin \omega t \tag{6.24}$$

Taking derivatives with respect to time yields

$$\frac{\partial \mathbf{E}}{\partial t} = -\omega \mathbf{E}^m \cos \omega t = +j\omega \mathbf{E}^m \sin \omega t \tag{6.25}$$

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} = -\omega^2 \mathbf{E}^m \sin \omega t \tag{6.26}$$

$$\frac{\partial \mathbf{H}}{\partial t} = -\omega \mathbf{H}^m \cos \omega t = +j\omega \mathbf{H}^m \sin \omega t \tag{6.27}$$

$$\frac{\partial^2 \mathbf{H}}{\partial t^2} = -\omega^2 \mathbf{H}^m \sin \omega t \tag{6.28}$$

Substitution of these terms in the wave equation yields

$$(\nabla^2 - j\omega\mu\sigma + \omega^2\mu\epsilon)\mathbf{E}^m = 0 \tag{6.29}$$

$$(\nabla^2 - j\omega\mu\sigma + \omega^2\mu\epsilon)\mathbf{H}^m = 0 \tag{6.30}$$

Now, as a matter of definition, let

$$k^2 \triangleq \omega^2\mu\epsilon - j\omega\mu\sigma \tag{6.31}$$

The wave equations may then be written as

$$(\nabla^2 + k^2)\mathbf{E}^m = 0 \tag{6.32}$$

$$(\nabla^2 + k^2)\mathbf{H}^m = 0 \tag{6.33}$$

6.4 Characteristics of the Transverse Electric (TE) Mode

For the transverse electric (TE) mode of transmission, it will be recalled that the electric field intensity vector **E** had only an X component. Hence

$$\mathbf{E}^m = \mathbf{i}E_x^m \tag{6.33}$$

where **i** is the unit vector along the X axis. Since this field is parallel to the metal plates, and since the parallel component of an electric field must vanish at the surface of a perfect conductor, then it follows that E_x must vary as some function of the distance along the Y axis, being zero at $y = 0$ and $y = b$, but finite elsewhere. This means that during the transient period following the initial production of the field, a standing wave of electric field intensity **E** must be produced across the Y dimension of the waveguide, resulting from the propagation and reflection of traveling waves across the guide.

Substitute

$$\mathbf{E}^m = \mathbf{i}E_x^m \tag{6.34}$$

into the wave equation

$$(\nabla^2 + k^2)\mathbf{E}^m = 0 \tag{6.35}$$

Now, since the Laplacian is

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{6.36}$$

and since the conducting planes are of infinite extent along the X axis, then

$$\frac{\partial^2 \mathbf{E}^m}{\partial x^2} = 0 \tag{6.37}$$

and the wave equation reduces to

$$\frac{\partial^2 E_x^m}{\partial y^2} + \frac{\partial^2 E_x^m}{\partial z^2} + k^2 E_x^m = 0 \tag{6.38}$$

Now, assume that the electric field intensity E_x^m depends upon the distance z in the direction of propagation according to the relationship

$$E_x^m = E_x^o e^{-\gamma z} \tag{6.39}$$

where γ = propagation constant of the same form as that encountered

in transmission line theory. Substituting Eq. (6.39) into the wave equation as given by (6.38) yields

$$\frac{\partial^2 E_x{}^o}{\partial y^2} + (k^2 + \gamma^2)E_x{}^0 = 0 \tag{6.40}$$

This is a second-order, linear, differential equation which may be solved by any conventional method, yielding

$$E_x{}^o = A \sin\sqrt{k^2 + \gamma^2}\, y + B \cos\sqrt{k^2 + \gamma^2}\, y \tag{6.41}$$

The arbitrary constants A and B will, in general, be determined by the character of the generator voltage. However, the two known boundary conditions on the electric field intensity at the metal surfaces make it possible to relate the value of $E_x{}^o$ to the physical dimensions of the waveguide. These boundary conditions are

$$\left.\begin{array}{l} (1)\ \text{When } y = 0 \text{ then } E_x = 0 \\ (2)\ \text{When } y = b \text{ then } E_x = 0 \end{array}\right\} \tag{6.42}$$

Apply boundary condition (1) to Eq. (6.41), obtaining

$$0 = 0 + B$$

so

$$B = 0$$

Consequently,

$$E_x^{\bullet} = A \sin\sqrt{k^2 + \gamma^2}\, y \tag{6.43}$$

Now apply boundary condition (2), obtaining

$$0 = A \sin\sqrt{k^2 + \gamma^2}\, b \tag{6.44}$$

There are two possibilities: A can be zero, which would be trivial; or

$$\sin\sqrt{k^2 + \gamma^2}\, b = 0 \tag{6.45}$$

Equation (6.45) will be satisfied whenever

$$\sqrt{k^2 + \gamma^2}\, b = \text{an integral multiple of } \pi \tag{6.46}$$

That is,

$$\sqrt{k^2 + \gamma^2}\, b = n\pi \tag{6.47}$$

where n = any integer, or

$$\sqrt{k^2 + \gamma^2} = \frac{n\pi}{b} \tag{6.48}$$

Substituting Eq. (6.48) into Eq. (6.43) yields

$$E_x^o = A \sin\left(\frac{n\pi}{b}\right)y \tag{6.49}$$

Hence, the electric field intensity varies sinusoidally along the Y coordinate, as one would expect from a consideration of standing-wave phenomena. The boundary conditions are satisfied and a solution of the wave equation obtained whenever the distance b between the plates is an integral number of half cycles of the sine wave across the Y dimension of the waveguide. Note that this does *not* imply a multiple of the half wavelength of the signal being propagated, an occasional misconception which sometimes arises. A separate field configuration exists for each integral value of n, and n is usually referred to as the *mode number*. The complete designation of the TE mode is then TE_n. Two possible configurations of

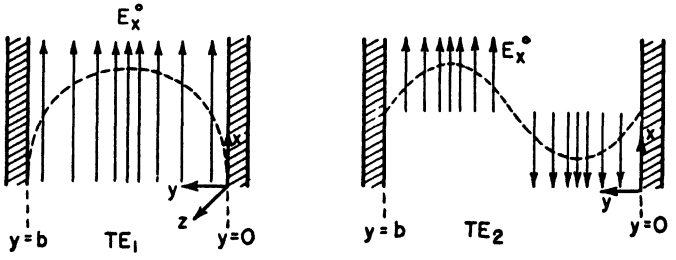


Fig. 6.6. Electric field intensity as a function of the distance y between the conducting plates for two different modes.

the electric field intensity are shown in Fig. 6.6. In general, the higher modes ($n > 1$) are avoided because effective methods of impedance matching at these higher modes have not been thoroughly developed up to the present time. The mode corresponding to $n = 1$ is commonly referred to as the *principal mode* for this reason. It is this mode, incidentally, which will propagate with the longest wavelength in the guide.

The complete expression for the electric field intensity may be obtained as follows:

$$E_x^o = A \sin\left(\frac{n\pi}{b}\right)y \qquad \text{by Eq. (6.49)}$$

but $$E_x^m = E_x^o e^{-\gamma z} \qquad \text{by Eq. (6.39)}$$

so $$E_x^m = Ae^{-\gamma z} \sin\left(\frac{n\pi}{b}\right)y \tag{6.50}$$

But $$E_x = E_x^m \sin \omega t \qquad \text{by Eq. (6.24)}$$

Hence, $$E_x = Ae^{-\gamma z} \sin\left(\frac{n\pi}{b}\right)y \sin \omega t \tag{6.51}$$

This is the complete expression for the electric field intensity in the dielectric of the waveguide, and it is worth noting that it is independent of the distance in the X direction, which was one of the original assumptions.

The magnetic field intensities corresponding to this electric field may be obtained from Eqs. (6.9) which are reproduced below and renumbered for convenience.

$$H_x^m = 0 \tag{6.52}$$

$$\frac{\partial E_x^m}{\partial z} = -j\omega\mu H_y^m \tag{6.53}$$

$$-\frac{\partial E_x^m}{\partial y} = -j\omega\mu H_z^m \tag{6.54}$$

Rewriting Eqs. (6.53) and (6.54)

$$H_y^m = -\frac{1}{j\omega\mu}\frac{\partial E_x^m}{\partial z} \tag{6.55}$$

$$H_z^m = -\frac{1}{j\omega\mu}\frac{\partial E_x^m}{\partial y} \tag{6.56}$$

Now, differentiate E_x^m, which is given by Eq. (6.50), partially with respect to z, and then repeat with respect to y. This process yields the following expressions:

$$\frac{\partial E_x^m}{\partial z} = -\gamma Ae^{-\gamma z} \sin\left(\frac{n\pi}{b}\right)y \tag{6.57}$$

$$\frac{\partial E_x^m}{\partial y} = \left(\frac{n\pi}{b}\right)Ae^{-\gamma z} \cos\left(\frac{n\pi}{b}\right)y \tag{6.58}$$

Substitute Eqs. (6.57) and (6.58) back into Eqs. (6.55) and (6.56) respectively, obtaining

$$H_y^m = \left(\frac{\gamma}{j\omega\mu}\right)Ae^{-\gamma z} \sin\left(\frac{n\pi}{b}\right)y = \left(\frac{\gamma}{j\omega\mu}\right)E_x^m \tag{6.59}$$

$$H_z^m = \left(\frac{n\pi}{j\omega\mu b}\right)Ae^{-\gamma z} \cos\left(\frac{n\pi}{b}\right)y \tag{6.60}$$

But, according to Eq. (6.24),

$$\mathbf{H} = \mathbf{H}^m \sin \omega t$$

so
$$H_y = -j\left(\frac{\gamma}{\omega\mu}\right)Ae^{-\gamma z} \sin\left(\frac{n\pi}{b}\right)y \sin \omega t \tag{6.61}$$

and
$$H_z = -j\left(\frac{n\pi}{\omega\mu b}\right)Ae^{-\gamma z} \cos\left(\frac{n\pi}{b}\right)y \sin \omega t \tag{6.62}$$

It was shown in Eq. (6.51) that

$$E_x = Ae^{-\gamma z} \sin\left(\frac{n\pi}{b}\right)y \sin \omega t \tag{6.63}$$

The last three equations specify the field configurations for any of the TE_n modes.

The variation in the Y component of the magnetic field intensity vector is sinusoidal, a necessary condition since the normal component of the magnetic field intensity must vanish at a perfectly conducting metal surface. On the other hand, the Z component varies consinusoidally as a function of y, indicating a current standing wave that is a maximum at the metal surfaces. Such a characteristic is to be expected since the tangential component of the magnetic field intensity is continuous at a metal surface. These points are noted to make it apparent that both components of the magnetic field could be obtained directly from the wave equation by proper substitution of the boundary conditions.

6.5 The Propagation Constant

The constant γ appearing in the equations of the preceding articles specifies the dependence of the magnitudes of the field intensities upon the distance traveled in the direction of propagation according to the relationships

$$\left.\begin{aligned}\mathbf{E} &= \mathbf{E}^m e^{-\gamma z} \\ \mathbf{H} &= \mathbf{H}^m e^{-\gamma z}\end{aligned}\right\} \tag{6.64}$$

Consequently, it fulfills the same function and has the same nature as the conventional propagation constant associated with transmission lines. In the most general case it will be composed of an attenuation term and a phase shift term. So, define the propagation constant for any mode n as

$$\gamma_n = \alpha_n + j\beta_n \tag{6.65}$$

where α_n is the attenuation constant for mode n and β_n is the phase shift constant for mode n. In Eq. (6.48) it was shown that

$$\sqrt{k_n^2 + \gamma_n^2} = \frac{n\pi}{b} \tag{6.66}$$

Squaring both sides yields

$$k_n^2 + \gamma_n^2 = \left(\frac{n\pi}{b}\right)^2 \tag{6.67}$$

Now solve this equation for γ_n

$$\gamma_n = \sqrt{\left(\frac{n\pi}{b}\right)^2 - k_n^2} \tag{6.68}$$

By Eq. (6.31)

$$k_n^2 = \omega^2\mu\epsilon - j\omega\mu\sigma$$

which, when substituted in Eq. (6.68), yields

$$\gamma_n = \sqrt{\left(\frac{n\pi}{b}\right)^2 - \omega^2\mu\epsilon + j\omega\mu\sigma} \tag{6.69}$$

In most waveguides the dielectric losses are usually negligible, so that without much loss in accuracy, it could be assumed that there were no dielectric losses and hence $\sigma = 0$. In any case, in general, for good dielectrics, $\omega\mu\sigma \ll \omega^2\mu\epsilon$. Under this assumption, the expression for the propagation constant reduces to

$$\gamma_n = \sqrt{\left(\frac{n\pi}{b}\right)^2 - \omega^2\mu\epsilon} \tag{6.70}$$

or, factoring out a (-1),

$$\gamma_n = j\sqrt{\omega^2\mu\epsilon - \left(\frac{n\pi}{b}\right)^2} \tag{6.71}$$

Comparing Eq. (6.71) with the assumed form for the propagation constant,

$$\gamma_n = \alpha_n + j\beta_n$$

and equating reals and imaginaries, it can be deduced that

$$\alpha_n = 0 \tag{6.72}$$

and

$$\beta_n = \sqrt{\omega^2\mu\epsilon - \left(\frac{n\pi}{b}\right)^2} \tag{6.73}$$

However, β_n is a real positive number, so that Eq. (6.73) is satisfied only as long as

$$\omega^2\mu\epsilon > \left(\frac{n\pi}{b}\right)^2 \tag{6.74}$$

This is commonly called the *condition for propagation* since propagation can occur only as long as the propagation constant contains a phase shift term. The frequency at which

$$\omega^2\mu\epsilon = \left(\frac{n\pi}{b}\right)^2 \tag{6.75}$$

is the frequency which separates the region of transmission from the region of attenuation only. It is customarily called the *cutoff frequency*. Hence, letting $f = f_0$,

$$\omega_0^2\mu\epsilon = \left(\frac{n\pi}{b}\right)^2 \tag{6.76}$$

Taking the square root of both sides and dividing through by 2π reduces the equation to

$$f_0 = \frac{n}{2b}\sqrt{\frac{1}{\mu\epsilon}} \tag{6.77}$$

The radical, however, is the equation for phase velocity, designated by the term v_1. So the expression for the cutoff frequency may be written as

$$f_0 = \frac{nv_1}{2b} \qquad \text{where} \qquad v_1 = \sqrt{\frac{1}{\mu\epsilon}} \tag{6.78}$$

Should the frequency fall below cutoff, then

$$\omega^2\mu\epsilon < \left(\frac{n\pi}{b}\right)^2$$

so that

$$\gamma_n = \sqrt{\left(\frac{n\pi}{b}\right)^2 - \omega^2\mu\epsilon}$$

which is a real number. Since the propagation constant then consists of a real, or attenuation, term only, it can be concluded that f_0 divides the region of transmission from the region of attenuation, causing the waveguide to act essentially like a high pass filter as illustrated in Fig. 6.7.

It should be understood that, although it is possible to excite the waveguide in the desired mode at frequencies below cutoff, no real propagation of energy takes place down the guide. The input impedance of the waveguide will be a pure reactance, and the fields that are excited in the guide will diminish exponentially as a function of the distance measured from the point of excitation. This effect has been used practically to obtain microwave attenuators. They are called *cutoff attenuators*.

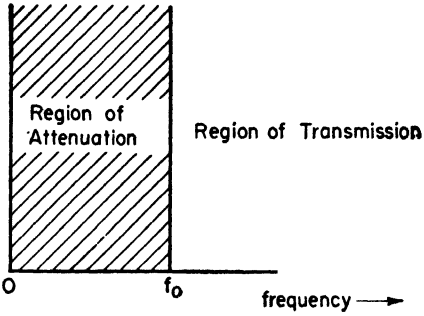


Fig. 6.7. High-pass filter action of a waveguide.

6.6 The TE_n Transmission Modes

In Art. 6.4 it was shown that, within the broad general classification of the TE mode, there are an infinite number of possible modes depending upon the magnitude of the integer n in the field intensity equations. These equations are rewritten below for convenience.

$$E_x = Ae^{-\gamma_n z} \sin\left(\frac{n\pi}{b}\right) y \sin \omega t$$

$$H_y = -j\left(\frac{\gamma n}{\omega\mu}\right) Ae^{-\gamma_n z} \sin\left(\frac{n\pi}{b}\right) y \sin \omega t$$

$$H_z = -j\left(\frac{n\pi}{\omega\mu b}\right) Ae^{-\gamma_n z} \cos\left(\frac{n\pi}{b}\right) y \sin \omega t$$

The *critical*, or cutoff, frequency is given by

$$f_0 = \frac{nv_1}{2b}$$

The cutoff wavelength may be found as follows:

$$\lambda_0 = \frac{v_1}{f_0} = \frac{2b}{n} \tag{6.79}$$

Solving for b yields

$$b = \tfrac{1}{2}\, n\lambda_0 \tag{6.80}$$

Equation (6.80) indicates that the dimension b of the waveguide is

critical and depends upon the particular mode which is to be excited. This supports the argument previously advanced based upon standing wave considerations.

According to Eq. (6.73)

$$\beta_n = \sqrt{\omega^2\mu\epsilon - \left(\frac{n\pi}{b}\right)^2}$$

but $$\frac{1}{\mu\epsilon} = v_1^2 = (\text{phase velocity})^2$$

Hence, $$\beta_n = \sqrt{\left(\frac{\omega}{v_1}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{6.81}$$

Now, as a matter of definition, let

$$\beta_n = \frac{2\pi}{\lambda_g} \tag{6.82}$$

Hence, $$\lambda_g = \frac{2\pi}{\beta_n} \tag{6.83}$$

Substituting for β_n yields

$$\lambda_g = \frac{2\pi}{\sqrt{(\omega/v_1)^2 - (n\pi/b)^2}} \tag{6.84}$$

For the principal mode (TE_1), $n = 1$, and the field equations reduce to

$$E_x = Ae^{-\gamma_1 z} \sin\left(\frac{\pi}{b}\right)y \sin \omega t$$

$$H_y = -j\left(\frac{\gamma_1}{\omega\mu}\right)Ae^{-\gamma_1 z} \sin\left(\frac{\pi}{b}\right)y \sin \omega t$$

$$H_z = -j\left(\frac{\pi}{\omega\mu b}\right)Ae^{-\gamma_1 z} \cos\left(\frac{\pi}{b}\right)y \sin \omega t$$

and the dimension b of the guide is

$$b = \tfrac{1}{2}\lambda_0 \tag{6.85}$$

The wavelength in the guide is

$$\lambda_g = \frac{2\pi}{\sqrt{(\omega/v_1)^2 - (\pi/b)^2}} = \frac{2\pi}{\beta_1} \tag{6.85}$$

The field distribution may be plotted as a function of the guide

dimensions by selecting any arbitrary instant of time and writing the field equations in the following form.

$$E_z = A_1 e^{-\gamma_1 z} \sin\left(\frac{\pi}{b}\right) y$$

$$H_y = A_2 e^{-\gamma_1 z} \sin\left(\frac{\pi}{b}\right) y$$

$$H_z = A_3 e^{-\gamma_1 z} \cos\left(\frac{\pi}{b}\right) y$$

where $\gamma_1 = j\beta_1$

and $\frac{H_y}{H_z} = \frac{\gamma_1 b}{\pi} \tan\left(\frac{\pi}{b}\right) y$

The results, when plotted, appear as shown in Fig. 6.8.

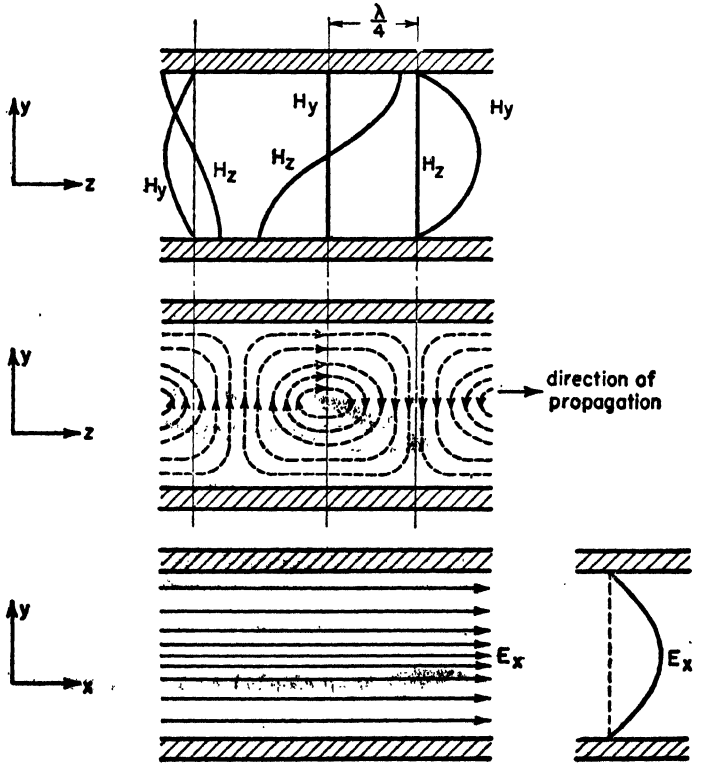


Fig. 6.8. Instantaneous field configurations for the TE_1 mode.

6.7 Resolution of the TE_1 Mode into Elementary Waves

Understanding of the phenomena occurring in waveguides, and the relation to transmission lines, is generally less difficult when the traveling wave in the waveguide is broken up into its constituent parts. This article will do this for the principal mode only since it is of the greatest importance. A similar analysis may be applied to the higher modes, but nothing of added significance would be revealed so that this single computation will suffice for present purposes.

For the TE_1 mode, the equation for the electric field intensity was shown to be

$$E_x = Ae^{-\gamma_1 z} \sin\left(\frac{\pi}{b}\right) y \sin \omega t \tag{6.87}$$

The sin $(\pi/b)y$ could be replaced by its exponential form as shown in Eq. (6.88).

$$E_x = Ae^{-\gamma_1 z}\left[\frac{e^{j\frac{\pi}{b}y} - e^{-j\frac{\pi}{b}y}}{2j}\right] \sin \omega t \tag{6.88}$$

Regrouping terms and replacing γ_1 by $j\beta_1$ yields

$$E_x = \frac{A}{2j}\left[e^{-j\left(\beta_1 z - \frac{\pi}{b}y\right)} - e^{-j\left(\beta_1 z + \frac{\pi}{b}y\right)}\right] \sin \omega t \tag{6.89}$$

Apparently the E_x wave is composed of two component waves, as evidenced by the presence of the two exponential terms, having equal amplitudes $(A/2j)\sin \omega t$, but traveling in different directions, the directions being given by the terms

$$\beta_1 z - \frac{\pi}{b} y$$

and

$$\beta_1 z + \frac{\pi}{b} y \tag{6.90}$$

Writing each component separately, and letting the common $(A/2j)\sin \omega t$ term be denoted by C gives

$$E_x^1 = Ce^{-j\left(\beta_1 z - \frac{\pi}{b}y\right)} \tag{6.91}$$

and

$$E_x^2 = Ce^{-j\left(\beta_1 z + \frac{\pi}{b}y\right)} \tag{6.92}$$

Therefore,

$$E_x = E_x^1 + E_x^2 \tag{6.93}$$

Consequently, the E_x^1 wave is traveling in the positive Z direction, but at an angle

$$\theta_1 = \tan^{-1}\left(-\frac{\pi}{b\beta_1}\right) \tag{6.94}$$

with the Z axis. Similarly, it can be deduced that the E_x^2 wave is traveling in the positive Z direction, but at an angle

$$\theta_2 = \tan^{-1}\left(+\frac{\pi}{b\beta_1}\right) \tag{6.95}$$

with the Z axis. It can be shown that each of these elementary waves is a plane wave and is therefore propagated with an energy velocity v_e numerically equal to the phase velocity. Figure 6.9

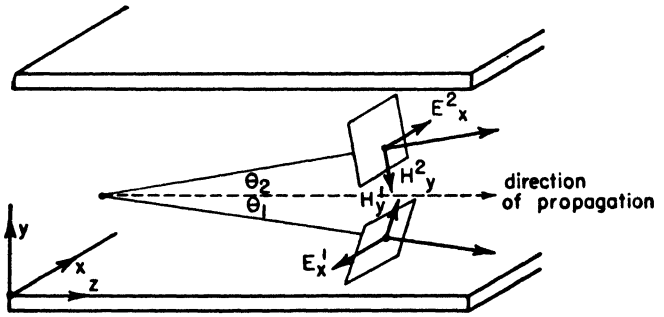


Fig. 6.9. Resolution of the TE_1 mode into two elementary waves.

shows the directions that the elementary waves take in their progression down the waveguide.

Thus, it is apparent that the TE_1 wave can be considered as a superposition of the two elementary plane waves; that is, the direction of propagation of each component is chosen in such a way that it is perpendicular to its component **E** and **H** fields. Hence, the TE_1 wave is a superposition of two plane waves which are reflected back and forth between the guiding metal plates.

The angles, θ_1 and θ_2, with which the elementary waves strike the guiding surface, depend upon the phase constant β_1, which in turn depends upon frequency. Equation (6.81) states

$$\beta_n = \sqrt{\left(\frac{\omega}{v_1}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{6.96}$$

Then, for the principal mode, since $n = 1$, this reduces to

$$\beta_1 = \sqrt{\left(\frac{\omega}{v_1}\right)^2 - \left(\frac{\pi}{b}\right)^2} = \sqrt{\left(\frac{2\pi f}{v_1}\right)^2 - \left(\frac{\pi}{b}\right)^2} \tag{6.97}$$

Factor out a π/b to obtain

$$\beta_1 = \frac{\pi}{b}\sqrt{f^2\left(\frac{2b}{v_1}\right)^2 - 1} \tag{6.98}$$

However, by Eq. (6.78), the cutoff frequency is

$$f_0 = \frac{nv_1}{2b} \tag{6.99}$$

which, for the principal mode, is

$$f_{0_1} = \frac{v_1}{2b} \tag{6.100}$$

Hence,

$$\beta_1 = \frac{\pi}{b}\sqrt{\left(\frac{f}{f_{0_1}}\right)^2 - 1} \tag{6.101}$$

Consequently, the expressions for the angles of incidence of the two elementary waves reduce to

$$\theta_1 = \tan^{-1} - \frac{1}{\sqrt{\left(\frac{f}{f_{0_1}}\right)^2 - 1}} \tag{6.102}$$

$$\theta_2 = \tan^{-1} + \frac{1}{\sqrt{\left(\frac{f}{f_{0_1}}\right)^2 - 1}} \tag{6.103}$$

Hence, as the operating frequency f approaches the critical or cutoff frequency f_{0_1}, the radical approaches zero. This means that both angles approach 90°, but with opposite signs. Hence, at the point where the two frequencies are equal, there is no velocity component of either of the elementary waves in the Z direction. No useful transmission occurs and the waves just bounce back and forth between the guiding plates. Since no energy is traveling down the guide, the *group* (energy) velocity is zero. If the operating frequency slightly exceeds the cutoff frequency, the two angles are not quite 90° so that a small velocity component in the Z direction does exist

and energy is delivered to the load at a comparatively slow velocity. As the frequency is made larger and larger, the angles become smaller and smaller, until, in the limiting case, the angles are both zero, the wavefront is then completely transverse to the direction of propagation, and the group velocity equals the phase velocity.

Consider Fig. 6.10 for a moment. This figure shows the crests and troughs of the elementary waves, that is, the locations of the

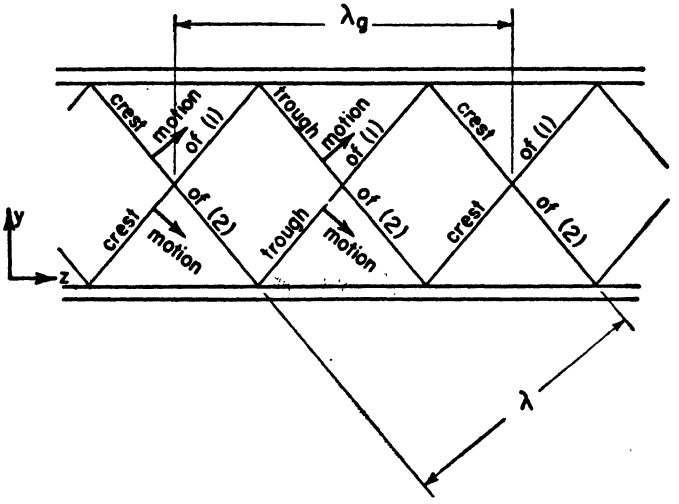


Fig. 6.10. Resolution of the TE_1 mode.

wavefronts. Propagation occurs in each case in a direction at right angles to these lines. The energy, then, is traveling back and forth between the conducting plates with a velocity equal to the velocity of light in that particular dielectric medium. The velocity of the energy in the Z direction is less than the velocity of light because of the zig-zag motion of the elementary waves. Thus it follows that the group (or energy) velocity is less than the velocity of light.

6.8 Phase and Group Velocity

Adequate understanding of the cutoff action of a waveguide demands a slight digression in order to clear up the definitions of the terms *phase velocity* and *group velocity*. Explanation by a simple analogy is generally sufficient.

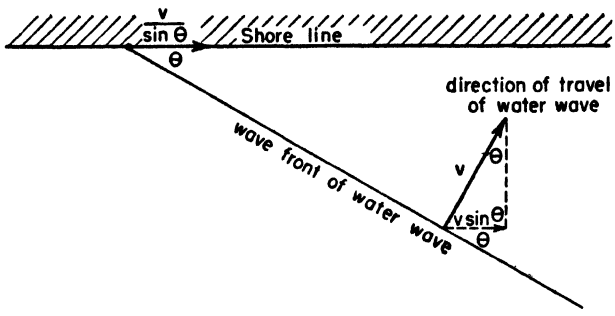


Fig. 6.11. Water wave analogy to explain phase and group velocity.

Consider a water wave approaching a straight shore line as shown in Fig. 6.11. Now, a water wave is a pure transverse wave so that its velocity perpendicular to the wavefront is the maximum energy velocity which it can attain (as an electromagnetic wave has the velocity of light as its maximum in any dielectric medium). The velocity of the energy in the wave, parallel to the shore line, is the component $v \sin \theta$. This, then, is designated as the group, or energy, velocity since energy transmission parallel to the shore line is comparable to the case of the waveguide previously discussed. Due to the presence of the sin θ term, the group velocity can never exceed the wave velocity v.

Now consider the intersection of the wave front with the shore line. As the wave rolls into the beach, this point, which is always at equiphase from one moment to another, moves with tremendous velocity. Its value is given by

$$v_1 = \frac{v}{\sin \theta}$$

which is always greater than the wave velocity v. Since it is the velocity of the equiphase points on the wave, it is appropriately called the phase velocity.

If the wave were approaching the shore line squarely, that is, if the wavefront were parallel to the shore line, the group velocity would be zero since no transmission of energy in the direction along the shore line occurs. On the other hand, the phase velocity must be infinite since all points on the wavefront intersect the shore line simultaneously.

Now consider the other extreme case, that is, when the wavefront is perpendicular to the shore line so that wave motion is entirely parallel to the beach. Then, both the phase and group velocities equal the wave velocity v.

The comparative relationship between these velocities is readily obtained since

$$v_g = v \sin \Theta = \text{group velocity}$$

$$v_1 = \frac{v}{\sin \Theta} = \text{phase velocity}$$

then $$v^2 = v_g v_1$$

Carrying the analogy over to the case of the elementary waves in the waveguide may be done very easily, and the cause of the cutoff characteristic is readily apparent as the case when the elementary waves have their wavefronts parallel to the guide surfaces so that the group velocity is zero and the phase velocity is infinite. In the waveguide case, the wave velocity designated by v in the preceding paragraphs is the velocity of light c. Hence,

$$c^2 = v_g v_1$$

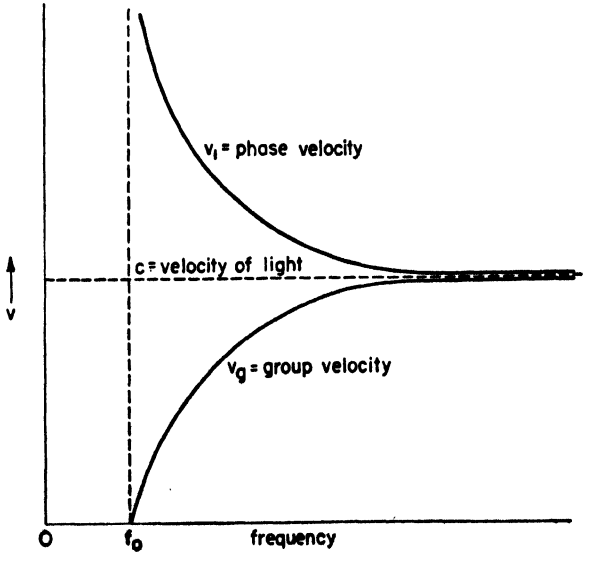


Fig. 6.12. Phase and group velocity as a function of frequency in a parallel plane waveguide.

Plots showing the interrelationships of these quantities appear in Fig. 6.12. The apparent paradox of a velocity greater than light is resolved when one considers that it is only the velocity of the equiphase points. A pulse, in a waveguide, would travel with the group velocity, not the phase velocity. In terms of the quantities shown in Fig. 6.10.

$$v_g = f\lambda$$
$$v_1 = f\lambda_g$$

Since $v_1 > v_g$, then $\lambda_g > \lambda$. Consequently, the *apparent* wavelength of the wave in the guide, λ_g, which is the one obtained by physical measurements, is always greater than the wavelength of the unguided wave.

In the extreme case where the phase and group velocities are equal, which occurs when the angles θ_1 and θ_2 are 0°, the wave is completely transverse and the two elementary waves merge into one single ray. Such a wave is called a transverse electromagnetic (TEM) wave since both the electric and magnetic field intensities are at right angles to the direction of propagation.

6.9 Wavelength in the Waveguide

In coaxial transmission lines the wavelength in the line always equals the wavelength associated with the velocity of light in that particular medium (assuming negligible attenuation) since the wave is TEM. In a waveguide, the wavelength in the guide always exceeds the wavelength of the unconfined wave except in the case of a TEM mode.

Define the following terms:

λ_g = wavelength in the guide = $\dfrac{2\pi}{\beta_n}$

λ = free-space wavelength = $\dfrac{c}{f}$

c = velocity of light = $\dfrac{1}{\sqrt{\mu_0\epsilon_0}}$

v_1 = phase velocity = $\dfrac{1}{\sqrt{\mu\epsilon}} = \dfrac{c}{\sqrt{K_e}}$

K_e = relative dielectric constant

Using Eq. (6.97) in the expression for λ_g above produces

$$\lambda_g = \frac{2\pi}{\sqrt{(2\pi f/v_1)^2 - (n\pi/b)^2}} \tag{6.104}$$

Rearrange terms as follows:

$$\lambda_g = \frac{v_1/f}{\sqrt{1 - (nv_1/2b)^2(1/f)^2}} = \frac{(v_1/f)}{\sqrt{1 - (f_0/f)^2}} \tag{6.105}$$

But,
$$f_0 = \frac{v_1}{\lambda_0} = \frac{1}{\sqrt{K_e}}\left(\frac{c}{\lambda_0}\right)$$

$$f = \frac{c}{\lambda}$$

$$\frac{v_1}{f} = \frac{1}{\sqrt{K_e}}\frac{c}{f} = \frac{\lambda}{\sqrt{K_e}}$$

Substituting these expressions into Eq. (6.105) yields

$$\lambda_g = \frac{\lambda}{\sqrt{K_e - (\lambda/\lambda_0)^2}} \tag{6.106}$$

Thus, the wavelength in the waveguide is always greater than the unguided wavelength. When the dielectric medium is air, the relative dielectric constant is 1, so

$$\lambda_g = \frac{\lambda}{\sqrt{1 - (\lambda/\lambda_0)^2}} \tag{6.107}$$

Thus, as the free-space wavelength approaches the cutoff wavelength,

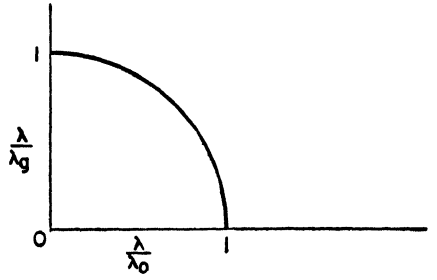


Fig. 6.13. Relationships between wavelength in the guide, cutoff wavelength, and free-space wavelength.

the guide wavelength approaches infinity. The expression given in Eq. (6.107) is plotted in Fig. 6.13. In this figure the ratio λ/λ_g is plotted against the ratio λ/λ_0 and the result is a quarter circle of unit radius.

6.10 Attenuation in a Waveguide Below Cutoff

As previously pointed out, if a possible mode of propagation is excited in a waveguide that is below cutoff, no real propagation of energy takes place for that mode. The two elementary waves just bounce back and forth between the walls of the guide, the wavefronts being parallel to the guide surfaces. Neglecting all losses, the input impedance will be a pure reactance and the magnitude of the fields excited will diminish exponentially from the point of excitation as given by the expression

$$\mathbf{E} = \mathbf{E}_0 e^{-\alpha z}$$

where α, the attenuation, is given by the expression following Eq. (6.78), that is

$$\alpha_n = \sqrt{\left(\frac{n\pi}{b}\right)^2 - \omega^2\mu\epsilon}$$

Assume the dielectric to be air. Rearrange terms as follows:

$$\alpha_n = \sqrt{\left(\frac{n\pi}{b}\right)^2 - \left(\frac{2\pi f}{c}\right)^2} = 2\pi\sqrt{\left(\frac{n}{2b}\right)^2 - \left(\frac{f}{c}\right)^2}$$

Substituting for the bracketed quantities yields

$$\alpha_n = 2\pi\sqrt{\left(\frac{1}{\lambda_0}\right)^2 - \left(\frac{1}{\lambda}\right)^2}$$

where λ is the free-space wavelength and λ_0 is the cutoff wavelength.

When the free-space wavelength is very much larger than the cutoff wavelength, the expression for the attenuation reduces to

$$\alpha_0 = \frac{2\pi}{\lambda_0}$$

In the perfectly ideal case of an infinitely long waveguide with perfectly conducting plates, no real propagation of energy takes place down the guide. However, it is possible to insert a probe or other suitable coupling device into the waveguide at a point some

distance away from the point of excitation and extract power. Hence, energy is being transferred down the guide. Since the energy removed by the pickup device is proportional to the square of the field intensity, which is itself a diminishing exponential function of the distance along the guide, a waveguide below cutoff can be used as a variable attenuator. The attenuation can be made very nearly a linear function of distance when $\lambda \gg \lambda_0$.

6.11 Propagation of the Principal TE Mode (TE_1)

The general condition for propagation as stated in Art. 6.6 is that

$$\omega^2\mu\epsilon > \left(\frac{n\pi}{b}\right)^2$$

or, in terms of the b dimension of the waveguide

$$b > \frac{n\pi}{\omega\sqrt{\mu\epsilon}}$$

This may also be written as

$$b > \frac{n}{2}\left(\frac{v_1}{f}\right)$$

Substituting for the bracket, and assuming that the dielectric in the waveguide is air, then

$$b > \frac{n\lambda}{2}$$

Hence, to excite the principal mode, $n = 1$ and

$$b > \frac{\lambda}{2}$$

However, to prevent the second mode, $n = 2$, from being propagated, the guide should be below cutoff for that mode, or

$$b < \lambda$$

Hence, for propagation in the principal mode *only*,

$$\frac{\lambda}{2} < b < \lambda$$

This is a very fortunate circumstance because it shows that the dimension of the waveguide is not hypercritical—it can fall within a

range of values, which, from the percentage viewpoint, is quite large. If this were not the case, a slight change in signal frequency would cause all sorts of undesirable effects ranging from cutting off transmission to exciting the wrong mode. Also, from the point of view of the manufacturer, the fact that b is not a *specific* size makes the tolerances very liberal. Economically, it allows production of waveguides for use over a rather wide *band* of frequencies, thus reducing manufacturing costs.

6.12 Skin Effect

The bulk of the discussion so far has proceeded upon the assumption that the metallic walls which provide the guiding action were perfect conductors. This allowed the boundary conditions to be specified at the metal surfaces. That is, it was stated that the tangential component of the electric field was zero at $y = 0$ and $y = b$. Actually, this ideal is never attained, but it can, by proper design, be closely approximated. However, a brief discussion of wave propagation in the metal walls is necessary at the outset.

If the walls of the waveguide are not perfect conductors, then a tangential component of the electric field exists at the air-metal interface. This field component couples the wave in the dielectric into the metal. The existence of the electric field intensity in the metal causes conduction currents to circulate that produce I^2R losses. Energy must then travel directly into the walls to supply these losses.

The primary purpose of the analysis that follows is to determine the extent to which these circulating currents penetrate into the metal walls. Using the coordinate system already specified in Art. 6.1, Fig. 6.3, the principal concern, then, is the variation in field intensity with respect to the variable y. Thus, select any arbitrary point (x,z) on one of the metal plates, and then investigate the effect on the field intensity as the point is moved into the metal parallel to the y-axis.

The wave equation is, according to Eq. (6.32),

$$(\nabla^2 + k^2)\mathbf{E} = 0$$

Assume transmission in the TE mode in the dielectric, that is

$$\mathbf{E} = \mathbf{i}E_z \qquad \text{and} \qquad E_y = E_x = 0$$

Only variations with respect to y are under observation, so

$$\nabla^2 = \frac{\partial^2}{\partial y^2}$$

Hence, the wave equation may be written as

$$\frac{\partial^2 E_x}{\partial y^2} + k^2 E_x = 0 \tag{6.108}$$

Solving this differential equation by any convenient method yields the following two possible solutions:

(1) $$E_x = E_0 e^{-jky}$$

(2) $$E_x = E_0 e^{+jky}$$

By ordinary logic it can be concluded that the first solution is the proper one since, as y approaches infinity, E_x certainly cannot approach infinity as given by the second solution. It would have to approach zero as specified by the first. Hence,

$$E_x = E_0 e^{-jky} \tag{6.109}$$

By the equation for the continuity of current

$$\mathbf{J}' = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$$

where $\mathbf{J}'$ is the total current density, $\mathbf{J}$ is the conduction current density, and $\partial \mathbf{D}/\partial t$ is the displacement current density. But, for a metal, the conduction current is so large that the displacement current is negligible in comparison. Hence,

$$\mathbf{J}' = \mathbf{J}$$

By Ohm's law

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

which may also be written as

$$\mathbf{J} = \sigma(\mathbf{E}_0 e^{-jky}) = \mathbf{J}_0 e^{-jky} \tag{6.110}$$

where $\mathbf{J}_0$ = surface current density $= \sigma \mathbf{E}_0$

The expression for the constant k was given in Eq. (6.31)as

$$k = \sqrt{\omega^2 \mu\epsilon - j\omega\mu\sigma}$$

This may be rewritten as

$$k = \sqrt{-j\omega\mu(+\sigma + j\omega\epsilon)}$$

But, for a metal, $\sigma \gg \omega\epsilon$, so

$$k = \sqrt{-j\omega\mu\sigma}$$

This may be put into a handier form by using the identity

$$(1 - j)^2 = -2j$$

When this is substituted in the equation for k, we get

$$jk = (1 + j)\sqrt{\frac{\omega\mu\sigma}{2}}$$

From a term-by-term comparison with the general form for the propagation constant, it can be concluded that

$$\alpha = \sqrt{\frac{\omega\mu\sigma}{2}} = \beta \tag{6.111}$$

Considering only the magnitude of the current density, then it is apparent that it diminishes exponentially as a function of y as shown by Fig. 6.14 and as given in the following equation:

$$J = J_0 e^{-\sqrt{(\omega\mu\sigma/2)}\, y} \tag{6.112}$$

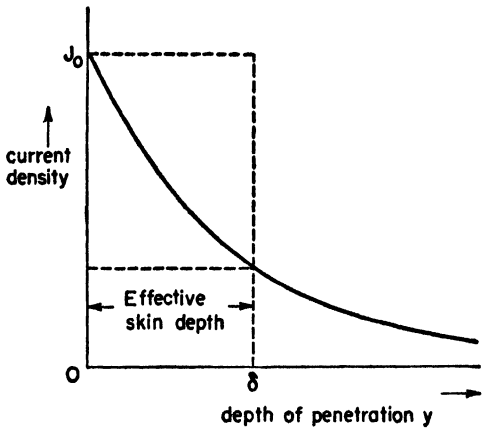


Fig. 6.14. Skin effect: depth of penetration of current.

The *current* can be evaluated by integrating the area under the curve for current density. That is,

$$I = \int_0^\infty J_0 e^{-\sqrt{(\omega\mu\sigma/2)}\, y}\, dy \tag{6.113}$$

Hence, $$I = J_0\sqrt{\frac{2}{\omega\mu\sigma}} \tag{6.114}$$

Let the radical in the above equation be designated as follows:

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} \qquad \text{therefore} \qquad I = J_0\delta \tag{6.115}$$

For many effects outside the metal, the same results can be obtained in calculations by assuming that the current is uniformly distributed over a strip of depth δ. Beyond this point, the current is assumed to be zero. δ is usually called the *skin depth* and is used as an index of the *effective depth of penetration* of the current. The effective skin depth decreases with frequency so that at 3000 mcps it is only about 0.00012 cm for copper. Consequently, from the practical standpoint, it is concluded that the current is all at the surface of the metal.

This very slight penetration enhances waveguide manufacture considerably. It allows the main body to be constructed of any metal with the required degree of cheapness, machinability, strength, and so on, without special regard to the electrical properties. Brass is very commonly used. Then, a very thin layer of gold or silver, both very nearly perfect conductors, can be plated on to the waveguide. Practically all of the current flows in this highly conducting surface film so that a very close approximation to the ideal condition is obtained.

6.13 Characteristics of the Transverse Magnetic (TM) Mode

By applying the same method to the TM mode as was used for the TE mode in Art. 6.4, the following equation will be obtained as a solution to the wave equation.

$$H_z = (B \sin\sqrt{\gamma^2 + k^2}\, y + C \cos\sqrt{\gamma^2 + k^2}\, y)\epsilon^{-\gamma z} \sin \omega t \tag{6.116}$$

According to Eq. (6.10-c)

$$-\frac{\partial H_z}{\partial y} = (\sigma + j\omega\epsilon)E_z \tag{6.117}$$

Rearranging terms and solving for E_z yields

$$E_z = -\left(\frac{1}{\sigma + j\omega\epsilon}\right)\frac{\partial H_z}{\partial y} \tag{6.118}$$

Carry out the indicated differentiation

$$E_z = -\frac{\sqrt{\gamma^2 + k^2}}{\sigma + j\omega\epsilon} \left(B \cos\sqrt{\gamma^2 + k^2}\, y - C \sin\sqrt{\gamma^2 + k^2}\, y\right) e^{-\gamma z} \sin \omega t \tag{6.119}$$

This component of the electric field is tangential to the metal plates so that, if the metal is assumed to be a perfect conductor, then the following boundary conditions apply:

(1) When $y = 0$ then $E_z = 0$
(2) When $y = b$ then $E_z = 0$

Application of these boundary conditions by the method used in Art. 6.4 for the TE mode reduces Eq. (6.119) to

$$E_z = \frac{(n\pi/b)}{\sigma + j\omega\epsilon} Ce^{-\gamma z} \sin\left(\frac{n\pi}{b}\right) y \sin \omega t \tag{6.120}$$

Consequently, working backward, the equation for H_x is

$$H_x = Ce^{-\gamma z} \cos\left(\frac{n\pi}{b}\right) y \sin \omega t \tag{6.121}$$

To solve for the y component of the electric field, use Eq. (6.10-b) for the TM mode, that is

$$E_y = \frac{1}{\sigma + j\omega\epsilon} \frac{\partial H_x}{\partial z} \tag{6.122}$$

Carrying out the differentiation yields

$$E_y = -\left(\frac{\gamma}{\sigma + j\omega\epsilon}\right) Ce^{-\gamma z} \cos\left(\frac{n\pi}{b}\right) y \sin \omega t \tag{6.123}$$

In a good dielectric medium the conductivity is negligible so that the field equations reduce to

$$H_x = Ce^{-\gamma z} \cos\left(\frac{n\pi}{b}\right) y \sin \omega t$$

$$E_y = j\left(\frac{\gamma}{\omega\epsilon}\right) Ce^{-\gamma z} \cos\left(\frac{n\pi}{b}\right) y \sin \omega t$$

$$E_z = -j\left(\frac{n\pi}{\omega\epsilon b}\right) Ce^{-\gamma z} \sin\left(\frac{n\pi}{b}\right) y \sin \omega t$$

In every case, the n refers to the mode number in the same manner as in the case of the TE mode.

The equations for the other significant quantities are derived by exactly the same methods used in the preceding articles on the TE mode and, to avoid repetitiousness, will not be carried through here.

6.14 The Transverse Electromagnetic (TEM) Mode

A special case of the equations derived in Art. 6.13 exists when the mode number n is zero. This reduces the field equations to the following forms:

$$H_x = Ce^{-\gamma z} \sin \omega t \tag{6.124}$$

$$E_y = j\left(\frac{\gamma}{\omega\epsilon}\right)Ce^{-\gamma z} \sin \omega t = j\left(\frac{\gamma}{\omega\epsilon}\right)H_x \tag{6.125}$$

$$E_z = 0 \tag{6.126}$$

Hence, both the electric and magnetic fields are at right angles to the direction of propagation, making both **E** and **H** transverse. Consequently, this is called the transverse electromagnetic (TEM) mode of transmission. Note that no standing waves are set up across the y co-ordinate of the waveguide, so that we would be led to believe that there is no cutoff characteristic associated with this mode. Such is the case since the cutoff frequency f_0 is

$$f_0 = \frac{nv_1}{2b}$$

and when the mode number is zero, the cutoff frequency is necessarily zero. Consequently, in the lossless case, *all* frequencies are transmitted without attenuation and the idealized high pass filter characteristic discussed in a preceding article is not obtained.

The case is interesting because it closely resembles that of a dissipationless transmission line in which the electric and magnetic fields are everywhere perpendicular to each other and to the direction of propagation. The characteristic impedance of such a line is a pure resistance, independent of frequency, so that energy of any frequency is transmitted with equal ease The parallel plane waveguide discussed here is merely the limiting case of the open wire transmission line as more and more conductors are added.

It should be noted that the same result is *not* obtained if the mode number in the TE mode equations is set equal to zero. Doing this would yield

$$E_x = 0 \qquad H_y = 0 \qquad H_z = 0$$

The result is entirely consistent with the picture one would expect from standing-wave considerations or from the boundary conditions.

6.15 Specific Wave Impedance

The characteristic impedance for a waveguide is not quite the same thing as the characteristic impedance of a transmission line. This is so because the transmission line is customarily designed to operate in the dominant mode only, which is a TEM wave. Waveguides are not operated in this mode, generally, but principally in the TE or TM modes.

The most generally used definition, at the present time, is the *specific wave impedance*, which is the ratio of the transverse electric field to the transverse magnetic field for a given mode in the waveguide. Hence, for the TE modes,

$$Z_w = \frac{E_x}{H_y} \tag{6.127}$$

where $$E_x = Ae^{-\gamma_n z} \sin\left(\frac{n\pi}{b}\right) y \sin \omega t$$

$$H_y = -j\left(\frac{\gamma n}{\omega\mu}\right) Ae^{-\gamma_n z} \sin\left(\frac{n\pi}{b}\right) y \sin \omega t$$

Substitution of these equations into Eq. (6.127) produces

$$Z_w = \frac{j\omega\mu}{\gamma_n}$$

In the lossless case there is no attenuation and

$$\gamma_n = j\beta_n$$

and the expression for the specific wave impedance becomes

$$Z_w = \frac{\omega\mu}{\beta_n}$$

However, by Eq. (6.82)

$$\lambda_g = \frac{2\pi}{\beta_n} \quad \text{or} \quad \beta_n = \frac{2\pi}{\lambda_g}$$

Substituting for β_n,

$$Z_w = \frac{\omega\mu\lambda_g}{2\pi} = \frac{2\pi f\mu\lambda_g}{2\pi} = f\mu\lambda_g$$

Multiply and divide by the velocity of light c for the dielectric medium.

$$Z_w = c\mu\left(\frac{f}{c}\right)\lambda_g = c\mu\,\frac{\lambda_g}{\lambda}$$

where λ = wavelength of the unguided wave

c = velocity of light $= \sqrt{\frac{1}{\mu\epsilon}} = \sqrt{\frac{1}{\mu_0\epsilon_0}}\sqrt{\frac{1}{K_mK_e}}$

μ = permeability $= \mu_0 K_m$

Substituting these relationships into the expression for the specific wave impedance gives

$$Z_w = \sqrt{\frac{K_m}{K_e}}\,377\,\frac{\lambda_g}{\lambda} \qquad \text{since} \qquad \sqrt{\frac{\mu_0}{\epsilon_0}} = 377 \text{ ohms}$$

where K_m is the relative permeability and K_e is the relative dielectric constant.

Following the same procedure for the TM modes yields

$$Z_w = \sqrt{\frac{K_m}{K_e}}\,377\,\frac{\lambda}{\lambda_g}$$

For the TEM mode the wavelength in the guide and the wavelength of the unguided wave are the same. Hence, the specific wave impedance is

$$Z_w = \sqrt{\frac{K_m}{K_e}}\,377$$

If the dielectric is free space, then the impedance is

$$Z_w = 377 \text{ ohms}$$

This is the *intrinsic impedance* of free space.

PROBLEMS

6.1 Two parallel copper sheets are placed 10 cm apart in air and excited in the TE_1 mode. Evaluate the propagation constant at frequencies of (a) 100 mcps, (b) 1000 mcps, (c) 10,000 mcps. Indicate whether or not propagation occurs.

6.2 Repeat Problem 6.1 for the TE_2 mode.

6.3 Calculate the skin depth in the copper walls of a waveguide operated in the TE_1 mode at (a) 100 mcps, (b) 1000 mcps, (c) 10,000 mcps.

6.4 The b dimension of a waveguide is 20 cm and an air dielectric is used. The guide is operated at a frequency of 1000 mcps. Calculate the cutoff frequency, the wavelength in the waveguide, and the propagation constant, assuming operation in the TE_1 mode.

6.5 What happens to the factors just calculated in Problem 6.4 if a dielectric wax having a relative dielectric constant of 4 is substituted for the air?

6.6 Calculate the specific wave impedance for the two waveguides of Problems 6.4 and 6.5.

6.7 Calculate the angles θ_1 and θ_2 that the component waves of the TE_1 mode of Problem 6.4 make with the Z axis. Explain how this is related to the fact that $\lambda_g > \lambda$.

CHAPTER 7

WAVEGUIDES AND CAVITY RESONATORS

THIS CHAPTER is concerned with the extension of the ideas and techniques developed in Chap. 6 for the hypothetical parallel plane waveguide to the case of the practical waveguide and cavity resonator. The discussion is by no means complete, but the essentials are clearly set forth and the details of the different practical variations on the fundamental structures that are covered may be picked up from any of the standard handbooks detailing specific design applications.*

7.1 Modes of Transmission in Rectangular Pipes

Consider the rectangular pipe, oriented with respect to an assumed set of co-ordinates, as shown in Fig. 7.1. Make the following assumptions:

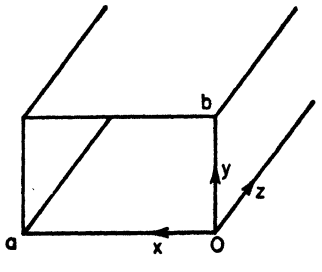


Fig. 7.1. Orientation of the waveguide with respect to an assumed set of coordinate axes.

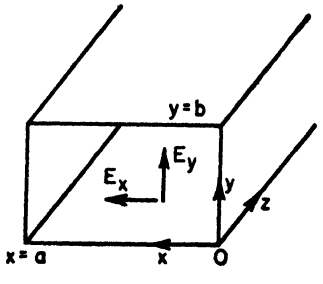


FIG. 7.2.

* For the qualitative material presented on cavities, the author drew rather heavily from the work of W. W. Hansen, as presented in *Microwave Transmission Design Data*, Sperry Gyroscope Co. publication 23-80, 1944. The author feels that this presentation touches the pertinent phenomena, and the considerations thereof, without bogging down into superfluous and repetitious mathematical development.

(1) The walls of the pipe are perfectly conducting.
(2) Propagation is along the Z axis.
(3) The wave is a harmonic function of time.

Under these conditions, it was shown in Art. 6.3 of the preceding chapter that transmission will occur if the wave equation is satisfied. The wave equation was derived and put into the forms shown by Eqs. (7.1) and (7.2).

$$(\nabla^2 + k^2)\mathbf{E}^m = 0 \tag{7.1}$$

$$(\nabla^2 + k^2)\mathbf{H}^m = 0 \tag{7.2}$$

The ∇^2 operator was previously defined as

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{7.3}$$

and the factor k was determined to be

$$k^2 = \omega^2\mu\epsilon - j\omega\mu\sigma \tag{7.4}$$

In these wave equations, the $\mathbf{E}^m$ and $\mathbf{H}^m$ terms are the maximum values of the harmonic time variations according to the defining relations

$$\mathbf{E} = \mathbf{E}^m \sin \omega t \tag{7.5}$$

$$\mathbf{H} = \mathbf{H}^m \sin \omega t \tag{7.6}$$

By assumption, and of necessity, steady-state propagation is in the Z direction. Thus, the wave is related to the direction of propagation according to the expression

$$\mathbf{E}^m = \mathbf{E}^0 e^{-\gamma z} \quad \text{and} \quad \mathbf{H}^m = H^0 e^{-\gamma z} \tag{7.7}$$

where $\gamma = \alpha + j\beta$ (propagation constant), α being the attenuation constant and β, the wavelength constant. Substitution of this relationship into the wave equation for the electric field yields

$$\frac{\partial^2 \mathbf{E}^0}{\partial x^2} + \frac{\partial^2 \mathbf{E}^0}{\partial y^2} + (k^2 + \gamma^2)\mathbf{E}^0 = 0 \tag{7.8}$$

and, for the magnetic field

$$\frac{\partial^2 \mathbf{H}^0}{\partial x^2} + \frac{\partial^2 \mathbf{H}^0}{\partial y^2} + (k^2 + \gamma^2)\mathbf{H}^0 = 0 \tag{7.9}$$

As in the case of the parallel plane waveguide, two distinct modal configurations are possible depending upon which field, electric or

magnetic, is made transverse to the direction of propagation. These two cases are:

(1) When the electric field has X and Y components only so that the electric field intensity vector lies entirely within the X-Y plane, perpendicular to the direction of propagation (TE mode).
(2) When the magnetic field has X and Y components only so that the magnetic-intensity vector is everywhere contained in the X-Y plane (TM mode).

Under these assumed field orientations, the second derivatives of both field intensity vector components must exist due to the boundary conditions imposed by the presence of the metal walls along both X and Y coordinates. As a consequence, the wave equation is a partial differential equation in the two variables X and Y, a fact which renders the solution just half as easy to obtain as it was in the case of the parallel plane waveguide where the differential equation contained only one variable. The most straightforward method of obtaining a solution is by the *method of separation of variables* which is briefly illustrated in the next article.

Actually, there are four variables, or dimensions, considered in the complete solution; x, y, z, and time t. However, the time variable was eliminated by assuming harmonic time dependence. The variable z was removed from the consideration by assuming that the wave was related to the direction of propagation according to an exponential term involving the propagation constant. This assumption was based upon previous experience with transmission lines. Thus, only the dependence upon the two variables x and y remains unspecified.

7.2 Method of Separation of Variables

Before proceeding to the solution of the wave equation, it is advisable to introduce the basic mathematical tool to be used, the method of separation of variables. Consider a partial differential equation in the two variables x and y of the form

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + C^2 f = 0 \tag{7.10}$$

If the second derivatives of the function with respect to each variable exist, the function f can be written implicitly as the product of

two other functions $X(x)$ and $Y(y)$ where $X(x)$ is the function of x only and $Y(y)$ is the function of y only. That is,

$$f(x,y) = X(x)Y(y) \tag{7.11}$$

or, omitting the arguments of each function,

$$f = XY \tag{7.12}$$

Taking the second derivatives of f with respect to each variable then gives

$$\frac{\partial^2 f}{\partial x^2} = Y\frac{d^2X}{dx^2} \tag{7.13}$$

$$\frac{\partial^2 f}{\partial y^2} = X\frac{d^2Y}{dy^2} \tag{7.14}$$

Substituting back into the given differential Eq. (7.10), results in the following expression:

$$Y\frac{d^2X}{dx^2} + X\frac{d^2Y}{dy^2} + XYC^2 = 0 \tag{7.15}$$

Now, divide through by the product of the two functions, XY, obtaining

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} + C^2 = 0 \tag{7.16}$$

Examine this equation carefully. It is evident that it is composed of three terms,

(1) One that is a function of x only $= \dfrac{1}{X}\dfrac{d^2X}{dx^2}$

(2) One that is a function of y only $= \dfrac{1}{Y}\dfrac{d^2Y}{dy^2}$

(3) A constant, independent of x and y, $= C^2$.

Thus, in words, Eq. (7.16) states that a function of x only plus a function of y only equals a constant, regardless of the values of x and y. This could be true *only* if the terms involving x and y are also constants. Hence, let

$$\frac{1}{X}\frac{d^2X}{dx^2} = K_1 = \text{constant} \tag{7.17}$$

$$\frac{1}{Y}\frac{d^2Y}{dy^2} = K_2 = \text{constant} \tag{7.18}$$

Substituting Eqs. (7.17) and (7.18) back into Eq. (7.16) yields the following two equations, each a function of a single variable.

$$\frac{1}{X}\frac{d^2X}{dx^2} + (K_2 + C^2) = 0 \tag{7.19}$$

$$\frac{1}{Y}\frac{d^2Y}{dy^2} + (K_1 + C^2) = 0 \tag{7.20}$$

The variables x and y have been separated. Rearranging terms yields

$$\frac{d^2X}{dx^2} + X(K_2 + C^2) = 0 \tag{7.21}$$

$$\frac{d^2Y}{dy^2} + Y(K_1 + C^2) = 0 \tag{7.22}$$

These are ordinary, second-order, linear differential equations with constant coefficients and have solutions, obtainable from any standard text on differential equations, of the following form:

$$X = A_1 \sin\sqrt{K_2+C^2}\,x + A_2 \cos\sqrt{K_2+C^2}\,x \tag{7.23}$$

$$Y = B_1 \sin\sqrt{K_1+C^2}\,y + B_2 \cos\sqrt{K_1+C^2}\,y \tag{7.24}$$

Consequently, since

$$f = XY$$

the product of the two functions yields four possible solutions:

$$f_1 = A_1 \sin\sqrt{K_2+C^2}\,x \cdot B_1 \sin\sqrt{K_1+C^2}\,y \tag{7.25}$$

$$f_2 = A_1 \sin\sqrt{K_2+C^2}\,x \cdot B_2 \cos\sqrt{K_1+C^2}\,y \tag{7.26}$$

$$f_3 = A_2 \cos\sqrt{K_2+C^2}\,x \cdot B_1 \sin\sqrt{K_1+C^2}\,y \tag{7.27}$$

$$f_4 = A_2 \cos\sqrt{K_2+C^2}\,x \cdot B_2 \cos\sqrt{K_1+C^2}\,y \tag{7.28}$$

The *correct* solution for any given problem is determined by substitution of specified boundary conditions.

7.3 Solution to the Wave Equation

Consider the TE mode. In this case the electric field intensity vector has only an x and a y component. That is,

$$\mathbf{E}^0 = \mathbf{i}E_x + \mathbf{j}E_y \tag{7.29}$$

$\mathbf{i}$ and $\mathbf{j}$ are the unit vectors in the x and y directions, respectively.

E_x and E_y are the magnitudes of the electric field intensity in these directions. Each component is itself a function of x and y, that is

$$E_x = E_x(x,y) \tag{7.30}$$

$$E_y = E_y(x,y) \tag{7.31}$$

and must satisfy the wave equation.

Consider the x component only. The wave equation is then

$$\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + (k^2 + \gamma^2)E_x = 0 \tag{7.32}$$

Using the method of separation of variables, assume that

$$E_x(x,y) = X(x)Y(y) \tag{7.33}$$

so that the wave equation reduces to

$$Y\frac{d^2X}{dx^2} + X\frac{d^2Y}{dy^2} + (k^2 + \gamma^2)XY = 0 \tag{7.34}$$

Divide the equation through by XY to obtain

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} + (k^2 + \gamma^2) = 0 \tag{7.35}$$

and the variables are separated.

It was shown in the preceding article that

$$\frac{1}{X}\frac{d^2X}{dx^2} = K_1 = \text{constant} \tag{7.36}$$

$$\frac{1}{Y}\frac{d^2Y}{dy^2} = K_2 = \text{constant} \tag{7.37}$$

These are unspecified, purely arbitrary constants. Consequently, the following two equations are obtained when these identities are substituted into the wave equation given by (7.35):

$$\frac{d^2X}{dx^2} + (K_2 + k^2 + \gamma^2)X = 0 \tag{7.38}$$

$$\frac{d^2Y}{dy^2} + (K_1 + k^2 + \gamma^2)\,Y = 0 \tag{7.39}$$

Letting

$$(K_1 + k^2 + \gamma^2) = c_1^2 \quad \text{and} \quad (K_2 + k^2 + \gamma^2) = c_2^2$$

then the equations have the following solutions:

$$X = A_1 \sin c_2 x + A_2 \cos c_2 x \tag{7.40}$$

$$Y = B_1 \sin c_1 y + B_2 \cos c_1 y \tag{7.41}$$

But since

$$E_z = XY \tag{7.42}$$

there are four possible solutions,

$$E_{z_1} = C_1' \sin c_2 x \sin c_1 y \qquad \text{where } C_1' = A_1 B_1 \tag{7.43}$$

$$E_{z_2} = C_2' \sin c_2 x \cos c_1 y \qquad \text{where } C_2' = A_1 B_2 \tag{7.44}$$

$$E_{z_3} = C_3' \cos c_2 x \sin c_1 y \qquad \text{where } C_3' = A_2 B_1 \tag{7.45}$$

$$E_{z_4} = C_4' \cos c_2 x \cos c_1 y \qquad \text{where } C_4' = A_2 B_2 \tag{7.46}$$

The proper solution and the unknown constants c_1 and c_2 are determined by the boundary conditions. The remaining constant, C', is determined by the generator.

Precisely the same procedure is followed to determine the possible solutions of the y component of the electric field intensity, and, for that matter, all of the components of the magnetic field intensity. For E_y, we obtain

$$E_{y_1} = D_1' \sin c_2 x \sin c_1 y \tag{7.47a}$$

$$E_{y_2} = D_2' \sin c_2 x \cos c_1 y \tag{7.47b}$$

$$E_{y_3} = D_3' \cos c_2 x \sin c_1 y \tag{7.47c}$$

$$E_{y_4} = D_4' \cos c_2 x \cos c_1 y \tag{7.47d}$$

These are the same constants c_1 and c_2 that occur in the solutions for E_z, because, in the separation of variables, in every case it is necessary to pick two *arbitrary* constants K_1 and K_2 so that

$$K_1 + K_2 + (k^2 + \gamma^2) = 0 \tag{7.48}$$

Consequently, since they are purely arbitrary, the same constants are used in all solutions of the component field intensities.

7.4 Substitution of the Boundary Conditions

For a complete and correct solution, the wave in the dielectric of the waveguide must satisfy

(1) The wave equation.
(2) The boundary conditions.
(3) The relationship that $\nabla \cdot \mathbf{E} = 0$.

The work thus far has yielded four solutions of the wave equation for each component of the electric field intensity that may or may not satisfy the remaining two conditions listed above.

Assuming the guiding walls to be perfect conductors, then the tangential components of the electric field intensity must vanish at the surface. In effect, this states that the voltage across a perfect short circuit must be zero. Thus, for the x and y components of the electric field intensity, it is possible to state the following boundary conditions

$$E_x = 0 \quad \text{when } y = 0 \quad \text{(boundary condition 1)}$$
$$E_x = 0 \quad \text{when } y = b \quad \text{(boundary condition 2)}$$
$$E_y = 0 \quad \text{when } x = 0 \quad \text{(boundary condition 3)}$$
$$E_y = 0 \quad \text{when } x = a \quad \text{(boundary condition 4)}$$

By the method of separation of variables, four possible solutions of the wave equation for E_x and E_y were found as given by Eqs. (7.43) through (7.47). Substitute boundary condition 1 into the solutions for E_x and boundary condition 3 into the solutions for E_y. This process yields

For E_x at $y = 0$	For E_y at $x = 0$
$E_{x1} = 0$	$E_{y1} = 0$
$E_{x2} = C_2' \sin c_2 x$	$E_{y2} = 0$
$E_{x3} = 0$	$E_{y3} = D_3' \sin c_1 y$
$E_{x4} = C_4' \cos c_2 x$	$E_{y4} = D_4' \cos c_1 y$

Thus, neglecting the trivial solutions obtainable when C_2', C_4', D_3', and D_4' are zero, it is apparent that only the solutions corresponding to E_{x1}, E_{x3}, E_{y1}, and E_{y2} meet the specified boundary conditions.

Now apply boundary condition 2 to the remaining solutions for E_x, and boundary condition 4 to those remaining for E_y.

For E_x at $y = b$	For E_y at $x = a$
$E_{x1} = 0 = C_1' \sin c_2 x \sin c_1 b$	$E_{y1} = D_1' \sin c_2 a \sin c_1 y = 0$
$E_{x3} = 0 = C_3' \cos c_2 x \sin c_1 b$	$E_{y2} = D_2' \sin c_2 a \cos c_1 y = 0$

This condition on E_x can be met for all values of x, neglecting trivial considerations of $C_1' = C_3' = 0$, only if

$$\sin c_1 b = 0$$

or, more specifically, only if

$$c_1 b = n\pi \qquad \text{where } n = \text{any integer, } 0, 1, 2, 3, \ldots$$

Solve for c_1 to obtain

$$c_1 = \frac{n\pi}{b}$$

Following the same procedure for E_y yields

$$c_2 = \frac{m\pi}{a} \qquad \text{where } m = \text{any integer, } 0, 1, 2, 3, \ldots$$

The possible solutions for E_x and E_y can now be written as

$$E_{x1} = C_1' \sin\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y \qquad E_{y1} = D_1' \sin\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y$$

$$E_{x2} = C_2' \cos\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y \qquad E_{y2} = D_2' \sin\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y$$

Condition (3), that $\nabla \cdot \mathbf{E} = 0$ for the correct solution, was given at the beginning of this article and states that the divergence of the electric field intensity is zero. More simply, it states that the rate of change of the vector $\mathbf{E}$, in the direction of the vector, must be zero. That is,

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} = 0$$

since $E_z = 0$ for the TE mode. It is evident that the divergence will be unequal to zero only when there is a starting or stopping of flux lines, that is, divergence can occur only in a region where flux lines originate or terminate. Consequently, for the ideal dielectric assumed for the waveguide, the divergence must be zero.

Take the derivative of each component as shown below:

$$\frac{\partial E_{x1}}{\partial x} = C_1'' \cos\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y$$

$$\frac{\partial E_{x2}}{\partial x} = C_2'' \sin\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y$$

$$\frac{\partial E_{y1}}{\partial y} = D_1'' \sin\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y$$

$$\frac{\partial E_{y2}}{\partial y} = D_2'' \sin\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y$$

Apply the condition that the divergence of $\mathbf{E}$ is zero by adding

$\partial E_{x1}/\partial x$ and $\partial E_{y1}/\partial y$ and equating it to zero. This yields a sum that can be zero for all values of x, y, m, and n only if

$$C_1'' = D_1'' = 0$$

However, summing the two derivatives $\partial E_{x2}/\partial x$ and $\partial E_{y2}/\partial y$ gives

$$\sin\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y\ (C_3'' + D_2'') = 0$$

This *can* be zero for all values of x, y, m, and n, if $C_3'' = -D_2''$. Since this combination satisfies the condition that the divergence of the electric field intensity be zero, then the only solutions that fit all three specified requirements are

$$E_x = E_{x3} = C_1 \cos\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y \tag{7.49}$$

$$E_y = E_{y2} = C_2 \sin\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \tag{7.50}$$

where C_1 and C_2 are new constants replacing those previously used.

The particular modal configuration in the waveguide is specified by the values of the integers m and n. Thus, transmission is said to be in the $TE_{m,n}$ modes.

The three components of the magnetic field intensity vector may be determined in the same manner, using the wave equation, the boundary conditions for the magnetic field intensity at the surface of a perfect conductor, and the divergence equation $\nabla \cdot \mathbf{H} = 0$. The solutions obtained are

$$H_x = D_1 \sin\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \tag{7.51}$$

$$H_y = D_2 \cos\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y \tag{7.52}$$

$$H_z = D \cos\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \tag{7.53}$$

where D_1, D_2, and D are new constants different from those previously used with E_y.

The five constants C_1, C_2, D_1, D_2, and D are all interrelated in such a way that only one real unspecified constant exists. It is determined by the generator. The interrelationships are best calculated from Maxwell's equations by the method outlined in the next article.

7.5 The $TE_{m,n}$ Modes

It was shown in Art. 7.3 that

$$K_1 + K_2 + k^2 + \gamma^2 = 0$$

Now add and subtract the quantity $(k^2 + \gamma^2)$, and collect terms as follows:

$$(K_1 + k^2 + \gamma^2) + (K_2 + k^2 + \gamma^2) - (k^2 + \gamma^2) = 0$$

However, in Art. 7.4 it developed that

$$(K_1 + k^2 + \gamma^2) = c_1^2 = \left(\frac{n\pi}{b}\right)^2$$

$$(K_2 + k^2 + \gamma^2) = c_2^2 = \left(\frac{m\pi}{a}\right)^2$$

Hence, substituting and solving for γ^2 yields

$$\gamma_{m,n}^2 = -k^2 + \left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right] \tag{7.54}$$

But, according to Eq. (7.4)

$$k^2 = \omega^2\mu\epsilon - j\omega\mu\sigma = \frac{\omega^2}{c^2} - j\omega\mu\sigma$$

Assuming a lossless dielectric, $\sigma = 0$, then

$$k^2 = \frac{\omega^2}{c^2}$$

and the expression for the propagation constant $\gamma_{m,n}^2$ becomes

$$\gamma_{m,n}^2 = -\left\{\frac{\omega^2}{c^2} - \left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right]\right\}$$

Taking the square root of both sides yields

$$\gamma_{m,n} = j\beta_{m,n} = j\sqrt{\frac{\omega^2}{c^2} - \pi^2\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)} \tag{7.55}$$

This equation defines the phase constant for any given TE modal excitation.

Equations (7.49) to (7.53), derived in the preceding article for the components of the electric and magnetic field intensities, define the $TE_{m,n}$ modes. When the time dependence specified in Eqs.

(7.5) and (7.6) are inserted, and when the dependence upon z given in Eq. (7.7) is included, the modal equations become

$$E_x = C_1 \cos\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta_{m,n}z) \tag{7.56}$$

$$E_y = C_2 \sin\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta_{m,n}z) \tag{7.57}$$

$$E_z = 0 \tag{7.58}$$

$$H_x = D_1 \sin\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta_{m,n}z) \tag{7.59}$$

$$H_y = D_2 \cos\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta_{m,n}z) \tag{7.60}$$

$$H_z = D \cos\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta_{m,n}z) \tag{7.61}$$

where $$e^{j\beta z} \sin \omega t = \sin(\omega t - \beta z)$$

Since the dielectric has been assumed to be lossless, Maxwell's equations are

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t} \qquad \text{and} \qquad \nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}$$

Thus, in component form, and recalling that $E_z = 0$ for the $TE_{m,n}$ modes, the field equations become

$$\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = \epsilon \frac{\partial E_x}{\partial t} \tag{7.62a}$$

$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = \epsilon \frac{\partial E_y}{\partial t} \tag{7.62b}$$

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = 0 \tag{7.62c}$$

$$-\frac{\partial E_y}{\partial z} = -\mu \frac{\partial H_x}{\partial t} \tag{7.62d}$$

$$\frac{\partial E_x}{\partial z} = -\mu \frac{\partial H_y}{\partial t} \tag{7.62e}$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -\mu \frac{\partial H_z}{\partial t} \tag{7.62f}$$

Substitute the correct expressions for the field intensity components into Eqs. (7.62). Carry out the required derivatives and cancel like terms on either side of the equalities. This process yields

for (7.62a) $$-\left(\frac{n\pi}{b}\right)D + j\beta D_2 = j\omega\epsilon C_1 \tag{7.63a}$$

for (7.62b) $$-j\beta D_1 + \left(\frac{m\pi}{a}\right)D = j\omega\epsilon C_2 \tag{7.63b}$$

for (7.62d) $$\beta C_2 = -\omega\mu D_1 \tag{7.63c}$$

for (7.62e) $$\beta C_1 = \omega\mu D_2 \tag{7.63d}$$

where $\beta = \beta_{m,n}$ is understood.

Solving these equations simultaneously for C_1, C_2, D_1, and D_2 in terms of D yields

$$C_1 = \left(\frac{\omega\mu}{k^2}\right)\left(\frac{n\pi}{b}\right)D_0$$

$$C_2 = -\left(\frac{\omega\mu}{k^2}\right)\left(\frac{m\pi}{a}\right)D_0$$

$$D_1 = \left(\frac{\beta}{k^2}\right)\left(\frac{m\pi}{a}\right)D_0$$

$$D_2 = \left(\frac{\beta}{k^2}\right)\left(\frac{n\pi}{b}\right)D_0$$

where $D_0 = jD$ and $k^2 = \omega^2\mu\epsilon - \beta^2$

Hence, the complete equations for the $TE_{m,n}$ modes are

$$E_x = D_0\left(\frac{\omega\mu}{k^2}\right)\left(\frac{n\pi}{b}\right)\cos\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta z) \tag{7.64a}$$

$$E_y = -D_0\left(\frac{\omega\mu}{k^2}\right)\left(\frac{m\pi}{a}\right)\sin\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta z) \tag{7.64b}$$

$$E_z = 0 \tag{7.64c}$$

$$H_x = D_0\left(\frac{\beta}{k^2}\right)\left(\frac{m\pi}{a}\right)\sin\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta z) \tag{7.64d}$$

$$H_y = D_0\left(\frac{\beta}{k^2}\right)\left(\frac{n\pi}{b}\right)\cos\left(\frac{m\pi}{a}\right)x \sin\left(\frac{n\pi}{b}\right)y \sin(\omega t - \beta z) \tag{7.64e}$$

$$H_z = D_0 \cos\left(\frac{m\pi}{a}\right)x \cos\left(\frac{n\pi}{b}\right)y \cos(\omega t - \beta z) \tag{7.64f}$$

The specification of the $TE_{m,n}$ modes is now complete, excepting the one constant D_0 determined by the generator.

The *dominant mode* of transmission is the $TE_{1,0}$, where $m = 1$ and $n = 0$. The field equations then reduce to

$$E_x = 0$$

$$E_y = -D_0\left(\frac{\omega\mu}{k^2}\right)\left(\frac{\pi}{a}\right)\sin\left(\frac{\pi}{a}\right)x \sin(\omega t - \beta z)$$

$$E_z = 0$$

$$H_x = D_0\left(\frac{\beta}{k^2}\right)\left(\frac{\pi}{a}\right)\sin\left(\frac{\pi}{a}\right)x \sin(\omega t - \beta z)$$

$$H_y = 0$$

$$H_z = D_0 \cos\left(\frac{\pi}{a}\right)x \cos(\omega t - \beta z)$$

This will be recognized as the TE_1 mode encountered in the parallel plane waveguide.

7.6 The $TM_{m,n}$ Modes

If the procedure detailed in the preceding articles is followed for the $TM_{m,n}$ modes, the following equations result:

$$E_x = A_0\left(\frac{\beta}{k^2}\right)\left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi}{a}\right)x\sin\left(\frac{n\pi}{b}\right)y\sin(\omega t - \beta z) \tag{7.65a}$$

$$E_y = A_0\left(\frac{\beta}{k^2}\right)\left(\frac{n\pi}{b}\right)\sin\left(\frac{m\pi}{a}\right)x\cos\left(\frac{n\pi}{b}\right)y\sin(\omega t - \beta z) \tag{7.65b}$$

$$E_z = A_0\sin\left(\frac{m\pi}{a}\right)x\sin\left(\frac{n\pi}{b}\right)y\cos(\omega t - \beta z) \tag{7.65c}$$

$$H_x = A_0\left(\frac{\omega\epsilon}{k^2}\right)\left(\frac{n\pi}{b}\right)\sin\left(\frac{m\pi}{a}\right)x\cos\left(\frac{n\pi}{b}\right)y\sin(\omega t - \beta z) \tag{7.65d}$$

$$H_y = -A_0\left(\frac{\omega\epsilon}{k^2}\right)\left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi}{a}\right)x\sin\left(\frac{n\pi}{b}\right)y\sin(\omega t - \beta z) \tag{7.65e}$$

$$H_z = 0 \tag{7.65f}$$

where $$k^2 = \omega^2\mu\epsilon - \beta^2$$

Note the mathematical symmetry between these equations for the $TM_{m,n}$ modes and the ones for the $TE_{m,n}$ modes in the preceding article.

7.7 Field Distributions in the $TE_{m,n}$ Modes

The equations derived in Art. 7.5 may be plotted and, at any instant and for any given mode, the field distributions inside the waveguide may be determined. Typical plots are shown in Fig. 7.3 for the $TE_{1,0}$ mode, and in Fig. 7.4 for the $TE_{2,0}$ mode. A more complete case is sketched in Fig. 7.5 for the $TE_{1,1}$ mode. Any of the other higher modes may be plotted in the same manner. Generally speaking, however, only the dominant mode is currently used in practical work, largely because of impedance-matching difficulties with the higher modes.

7.8 Cutoff

The expression for the phase constant β was derived in Eq. (7.55) in the form

$$\beta = \sqrt{\frac{\omega^2}{c^2} - \pi^2\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)}$$

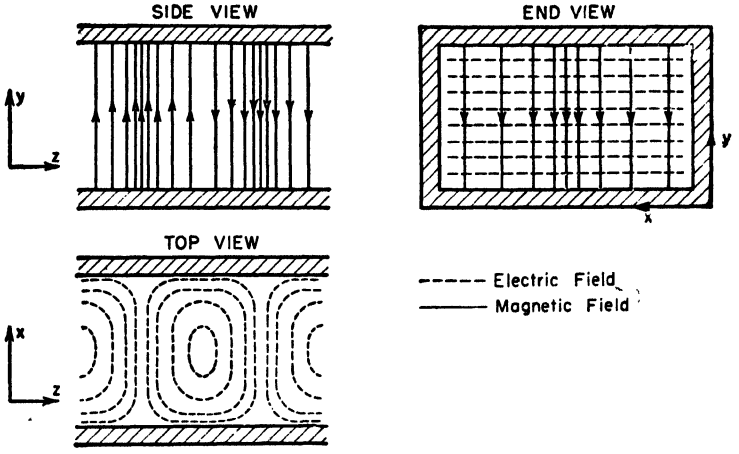


Fig. 7.3. $TE_{1,0}$ mode in rectangular waveguide.

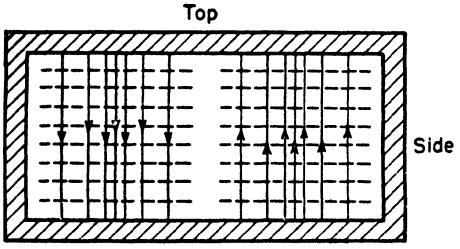


Fig. 7.4. $TE_{2,0}$ mode in rectangular waveguide.

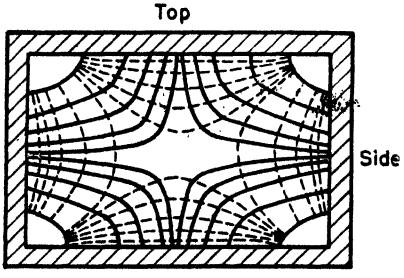


Fig. 7.5. $TE_{1,1}$ mode in rectangular waveguide.

But, $\omega = 2\pi f$ and $\frac{f}{c} = \frac{1}{\lambda}$

where c is the velocity of light. Hence,

$$\beta = \pi\sqrt{\frac{4}{\lambda^2} - \left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)}$$

At the cutoff wavelength, λ_c, no propagation occurs, so that $\beta = 0$. Hence, inserting this condition into the equation for the phase constant yields

$$\frac{4}{\lambda_c^2} - \left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right) = 0 \quad \text{or} \quad \lambda_c = \frac{2}{\sqrt{(m/a)^2 + (n/b)^2}} \tag{7.66}$$

This formula is true for either $TE_{m,n}$ or $TM_{m,n}$ modes. λ_c is the limiting wavelength that will propagate down the guide.

A rather interesting and informative graph may be obtained by plotting the size of the pipe necessary for propagation, using a/λ as the abscissa and b/λ as the ordinate. The result appears in Fig. 7.6. The small arrows indicate the boundaries between sizes where modes will propagate and where they will not. The arrows point toward the sizes that *do* propagate that mode. For example, the $TE_{2,0}$ mode will propagate in all sizes of waveguide for which a/λ is greater than 1. The $TE_{1,0}$ mode propagates in all pipes for which a/λ is greater than $\frac{1}{2}$. The $TE_{0,1}$ mode propagates when b/λ is greater than $\frac{1}{2}$, and so on. Only the dominant, or $TE_{1,0}$, mode is propagated if

$$\frac{1}{2} < \frac{a}{\lambda} < 1 \quad \text{and} \quad 0 < \frac{b}{\lambda} < \frac{1}{2}$$

It is also apparent from this chart that more than one mode can propagate at the same time unless the pipe dimensions are chosen in the proper way. The dimensions of the waveguide are selected so that operation is below cutoff for all undesired modes.

7.9 Power Transmitted and Characteristic Impedance

The Poynting vector **S** is defined as the vector product of the electric and magnetic field intensity vectors. That is,

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}$$

For the case of the dominant mode, since

$$E_x = E_z = H_y = 0$$

and if only the Z component of the Poynting vector is considered,

$$S_z = E_y H_x.$$

This product gives the instantaneous power transmitted, per unit

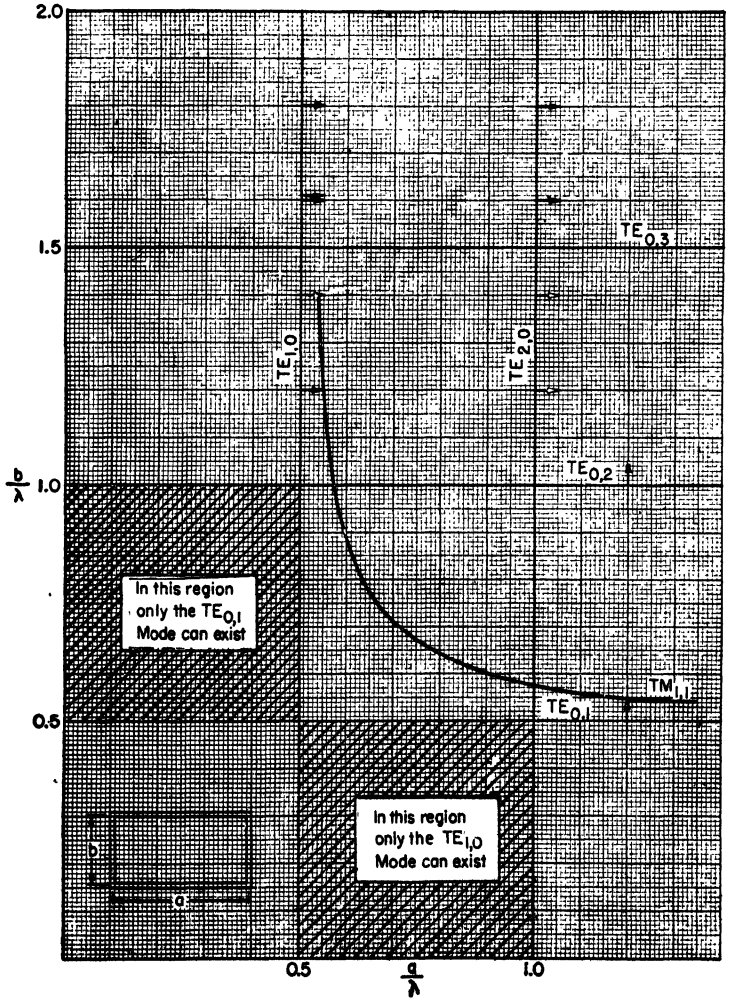


Fig. 7.6. Pipe size necessary for propagation of some of the lower modes in rectangular waveguide.

cross-sectional area, toward the load. A slight rearrangement of terms produces

$$S_z = \left(\frac{E_y}{H_x}\right) H_x^2 = Z_w H_x^2$$

where Z_w is the specific wave impedance. For the $TE_{1,0}$ mode, since

$$H_x = H_0 \sin\left(\frac{\pi}{a}\right) x \sin(\omega t - \beta z)$$

where H_0 is the peak value of the magnetic field intensity, the Z component of the Poynting vector is

$$S_z = Z_w H_0^2 \sin^2\left(\frac{\pi}{a}\right) x \sin^2(\omega t - \beta z)$$

The time average over one period τ is then

$$p_z = \frac{1}{\tau}\int_0^{\tau} S_z dt = \frac{1}{2\pi}\int_0^{2\pi} Z_w H_0^2 \sin^2\left(\frac{\pi}{a}\right) x \sin^2(\omega t - \beta z) d(\omega t)$$

Carrying out the integration and substituting limits yields

$$p_z = \frac{Z_w}{2} H_0^2 \sin^2\left(\frac{\pi}{a}\right) x$$

This gives the peak power flow in the direction of the load across any unit cross section. To find the total power flow P_z through the guide toward the load, it is necessary to integrate p_z over the cross section of the waveguide. That is,

$$P_z = \int_0^a \int_0^b p_z \, dy \, dx = \int_0^a \int_0^b \left(\frac{Z_w H_0^2}{2}\right) \sin^2\left(\frac{m\pi}{a}\right) x \, dy \, dx$$

Integrating and substituting limits yields a power

$$P_z = \left(\frac{Z_w H_0^2}{4}\right) ab \tag{7.67}$$

Another relationship for the $TE_{1,0}$ mode, that is of considerable practical use, may be derived from Eq. (7.67). The peak voltage across the guide opening is

$$V = E_y b$$

where E_y is the peak value of the electric field intensity and b is the y

dimension of the waveguide. Moreover, by definition, the specific wave impedance is

$$Z_w = \frac{E_y}{H_0}$$

Hence, solving for the magnetic field,

$$H_0 = \frac{E_y}{Z_w} = \frac{V}{Z_w b}$$

Substitute this expression for H_0 into Eq (7.67) for the total power flow.

$$P_z = \frac{1}{4} V\left[\frac{V}{Z_w}\left(\frac{a}{b}\right)\right] \tag{7.68}$$

In terms of an assumed voltage and current, using peak values, the power flow across a unit cross section would be

$$p = \frac{1}{2} VI$$

and the total power flow across the waveguide cross section would be

$$P = \frac{1}{4} VI \tag{7.69}$$

Hence, from a term-by-term comparison of Eqs. (7.68) and (7.69)

$$I = \frac{V}{Z_w}\left(\frac{a}{b}\right)$$

The input impedance of a waveguide operating in the $TE_{1,0}$ mode is then

$$Z_c = \frac{V}{I} = Z_w\left(\frac{b}{a}\right) = 120\pi\left(\frac{\lambda_g}{\lambda_a}\right)\left(\frac{b}{a}\right) \tag{7.70}$$

This derivation has assumed that no higher modes were present, and that there were no standing waves along the Z axis of the waveguide. Consequently, the impedance Z_c given in Eq. (7.70) must be the characteristic impedance of the waveguide in the dominant mode. This equation is applicable to cases such as the one shown in Fig. 7.7 where two waveguides of different dimensions are to be connected together without introducing any reflection. Assuming that the dimensions a and b of the first waveguide are given, then b' may be

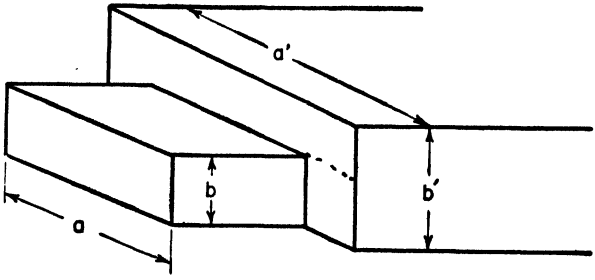


Fig. 7.7. Junction between waveguides of different sizes.

assumed for the second one. The cutoff wavelengths for both guides are then known, since

$$\lambda_c = 2a \qquad \text{and} \qquad \lambda_c' = 2a'$$

The wavelengths in the two guides are

$$\lambda_g = \frac{\lambda_a}{\sqrt{1 - \left(\frac{\lambda_a}{\lambda_c}\right)^2}} \qquad \text{and} \qquad \lambda_g' = \frac{\lambda_a}{\sqrt{1 - \left(\frac{\lambda_a}{\lambda_c'}\right)^2}}$$

This allows the calculation of the wave impedance for each guide as

$$Z_w = 120\pi \frac{\lambda_g}{\lambda_a} \qquad \text{and} \qquad Z_w' = 120\pi \frac{\lambda_g'}{\lambda_a}$$

If no reflection is to occur at the junction, the input impedances of both sections must be equal. That is,

$$Z = Z_w \frac{b}{a} = Z_w' \frac{b'}{a'}$$

and the unknown dimension a' is thereby determined to be

$$a' = \left(\frac{ab'}{b}\right)\frac{Z_w'}{Z_w}$$

7.10 Waveguide Excitation (Terminal Devices)

The excitation or reception of the various modes in rectangular pipes discussed so far may be accomplished by inserting probe antennas at points in the guide corresponding to points of maximum electric field intensity. Thus, the $TE_{1,0}$ mode is excited by a single probe located in the center of the X coordinate of the waveguide. The $TE_{2,0}$ mode would require two such probes spaced as indicated

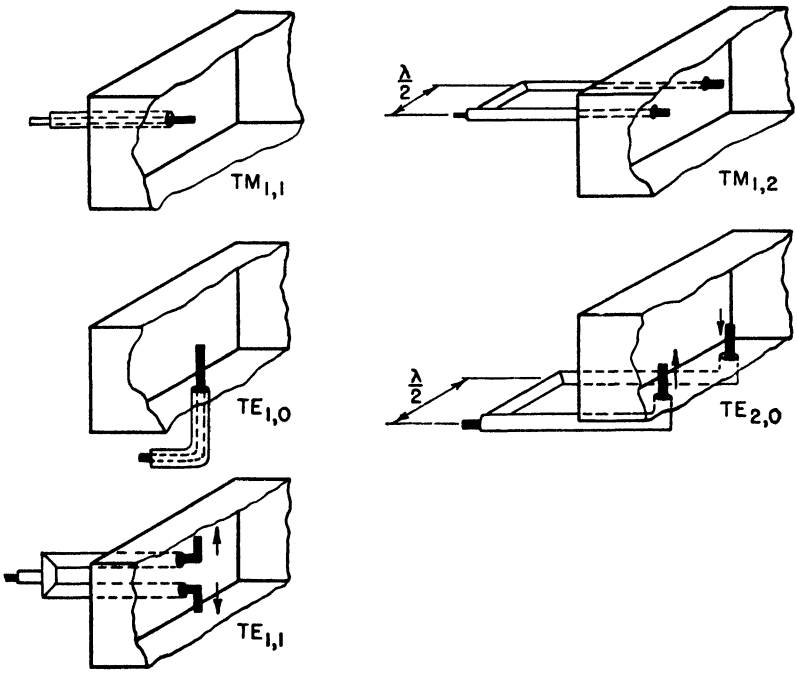


Fig. 7.8. Waveguide excitation (terminal devices) original with Chu and Barrow.

in Fig. 7.8. This figure also shows several other arrangements of terminal devices.

The probes are generally the extensions of the inner conductor of a coaxial cable. When more than one probe antenna is used, it is apparent from the field configurations of the various modes that the phase angle between the probes is important. The relative phases are indicated in Fig. 7.8 by arrows.

7.11 Impedance

In practical waveguide applications it is convenient to deal with impedances because it is generally easier to explain and to understand the operation of the various waveguide devices in such terms. Moreover, under certain conditions, the use of impedances makes it possible to represent the waveguide assembly as a transmission line. In turn, this permits the use of the circle diagrams discussed in Chap. 5.

Unfortunately, because the waveguide is a single-conductor system, it is not possible to define a unique current-voltage distribution analogous to that which occurs on a transmission line.

Two waveguide impedances have already been defined mathematically. The specific wave impedance, Z_w, was defined as the ratio of the transverse electric field to the transverse magnetic field. It was shown to be

$$Z_w = 120\pi\frac{\lambda_g}{\lambda_a} \qquad \text{for all TE modes}$$

$$Z_w = 120\pi\frac{\lambda_a}{\lambda_g} \qquad \text{for all TM modes}$$

This impedance is the same as the input impedance of an infinitely long line, or one terminated in its characteristic impedance, because the wave on a transmission line is normally a TEM wave which has *only* transverse electric and magnetic field components. Hence, if the waveguide could be excited in the TEM mode, the specific wave impedance would be the "characteristic impedance." However, since the waveguide is a single-conductor system, such excitation is physically impossible.

However, rectangular waveguides are ordinarily operated in the dominant mode. Assuming this operation, and no standing waves or higher modes, it was shown that the input impedance of the waveguide is given by

$$Z_c = Z_w\left(\frac{b}{a}\right)$$

This is analogous to the characteristic impedance for $TE_{1,0}$ mode operation.

If the waveguide is not terminated in an impedance equal to this "characteristic impedance," reflections occur, setting up standing waves along the axis of propagation of the waveguide. In this case, the ratio of transverse electric to transverse magnetic field is no longer a constant equal to the specific wave impedance, but varies in a complex manner with the length of the guide. It can be shown that at any point βl from the load that the input impedance of the waveguide is

$$Z = Z_c\frac{Z_L + jZ_c \tan \beta l}{Z_c + jZ_L \tan \beta l} \tag{7.71}$$

where Z_L is the ratio of the complex electric and magnetic fields at the

load and $Z_c = Z_w\ (b/a)$. The form of this equation is the same as that found for transmission lines and indicates that a transmission line can be used as the equivalent circuit of a waveguide if the following three facts are kept in mind:

(1) The impedances are ratios of field intensities, not of voltage to current.
(2) The wavelength on the equivalent transmission line is λ_g, not λ_a.
(3) The representation is valid for a single mode only; if higher modes are present, each such mode must be represented by a different equivalent line.

Consequently, if these conditions are recognized, the transmission line circle diagrams previously discussed may be used in the design of waveguide sections.

7.12 Discontinuities in Waveguides (Impedance Matching)

The two most commonly used types of waveguide discontinuities for impedance-matching purposes are (1) tuning screws and (2) windows. The tuning screw is simply a highly conducting metal screw inserted in the top or bottom wall of the waveguide as shown in Fig. 7.9. The window may take any one of very many forms, but the two most commonly used are the symmetrical windows shown in Fig. 7.10. They generally consist of thin, highly conducting, metal fins attached to the waveguide in a way that will reduce the cross-sectional area. Assuming the obstacles to be lossless, then they

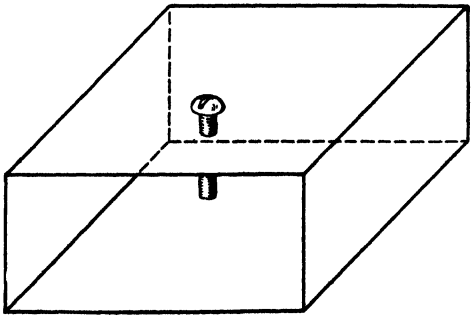


Fig. 7.9. Tuning screw.

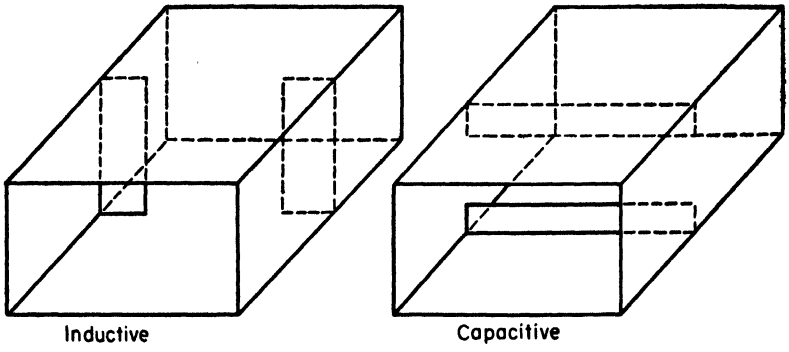


Fig. 7.10. Tuning windows.

introduce pure shunting susceptances (like stubs on transmission lines) into the waveguide. The equivalent circuits of these devices are shown in Fig. (7.11). Typical susceptance curves are shown in Figs. (7.12), (7.13), and (7.14).

The mathematical verification of these curves is difficult and exceedingly laborious. However, a fair qualitative understanding may be obtained from a consideration of the field configurations produced by the obstacles. When the incident wave strikes the obstruction, the higher modes are excited. They are necessary in order to satisfy the boundary conditions imposed by the introduction of the obstacles. However, even though they are generated they do not propagate—waveguide dimensions are such that only the principal mode propagates—but remain localized around the obstruction. The higher modes are the source of the effective shunting susceptances. If the higher modes developed are predominantly TM, the equivalent shunt susceptance is inductive; but, if TE modes predominate, it is capacitive.

The tuning screw behaves in much the same manner as an antenna if due consideration is accorded its images caused by the conducting plates which form the waveguide. The impedance of a dipole antenna varies in much the same fashion as an open-circuited transmission line, being capacitive when the half length of the antenna is less than a quarter of a wavelength, and inductive when greater than a quarter wavelength. Thus, we would be inclined to believe that when the insertion of the tuning screw is less than $\lambda/4$ that it introduces a

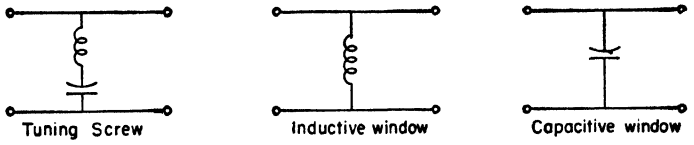


Fig. 7.11. Equivalent circuits for the matching elements of Figs. 7.9 and 7.10.

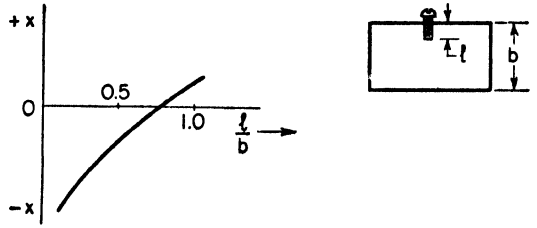


Fig. 7.12. Tuning-screw reactance curve.

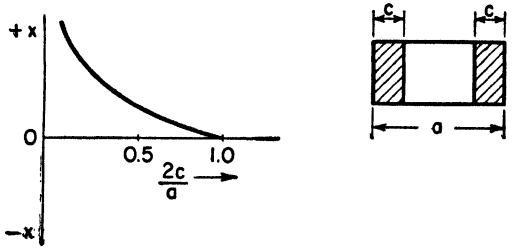


Fig. 7.13. Symmetrical inductive window-reactance curve.

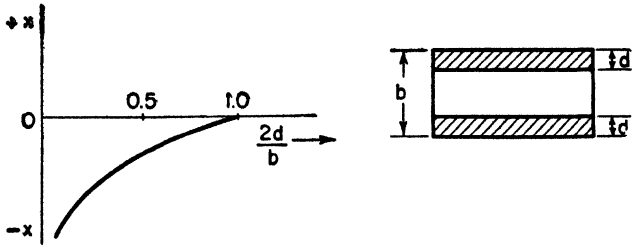


Fig. 7.14. Symmetrical capacitive window-reactance curve.

capacitive susceptance; and an inductive susceptance when greater than $\lambda/4$. This is approximately true.

An example of the practical application of this information should serve to clear up the many loose ends. Suppose that an inductive window is to be used to eliminate the standing waves on a waveguide connecting an oscillator to its antenna system, assuming operation in the $TE_{1,0}$ mode. The waveguide dimensions a and b are known, as are the oscillator wavelength λ_a and the input impedance of the antenna $Z_A = R_A + jX_A$. It is necessary to determine the location of the window and the susceptance required. The problem is recognizable as analogous to the single stub on a transmission line. The procedure for design is as follows.

Calculate the cutoff wavelength λ_c from

$$\lambda_c = 2a$$

and then the wavelength in the waveguide, using the following equation:

$$\lambda_g = \frac{\lambda_a}{\sqrt{1 - (\lambda_a/\lambda_c)^2}}$$

The characteristic impedance of the waveguide in the $TE_{1,0}$ mode may be found from the equation

$$Z_c = Z_w\left(\frac{b}{a}\right) = 120\pi\left(\frac{\lambda_g}{\lambda_a}\right)\left(\frac{b}{a}\right)$$

Now normalize the antenna impedance, using the same method as that employed with transmission lines. That is,

$$z_A = \frac{Z_A}{Z_c} = \frac{R_A}{Z_c} + j\frac{X_A}{Z_c} = r_A + jx_A$$

Enter the circle diagram with this value of chart impedance.Convert it to admittance by rotating 90° on the appropriate constant-ρ circle to the point

$$y_A = g_A - jb_A$$

From this point, rotate clockwise on the same constant-ρ circle to the point where it intersects the line

$$y = 1 + jb$$

The angle turned out in the rotation is βd where β is the phase constant

of the waveguide, d is the distance from antenna to window, and b, the window susceptance in chart units. Hence,

$$B = Y_c b = \text{total window susceptance}$$

The actual dimensions of the window would be determined from a graph of the form shown in Fig. 7.13 or 7.14.

7.13 Waveguide T's, Bends, Corners, and Twists

Frequently, junctions between waveguides are necessary. For the rectangular section, typical T-junctions for the $TE_{1,0}$ mode are shown in Fig. (7.15). They are identified by either of two designations, as follows:

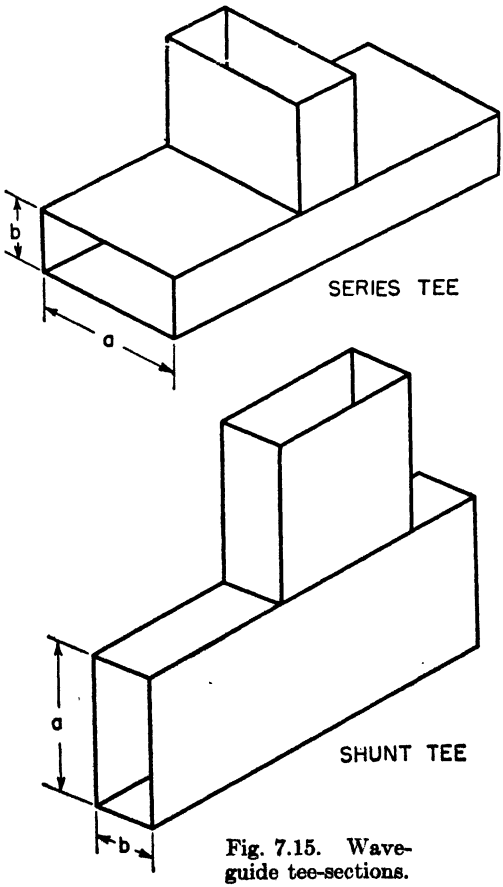


Fig. 7.15. Waveguide tee-sections.

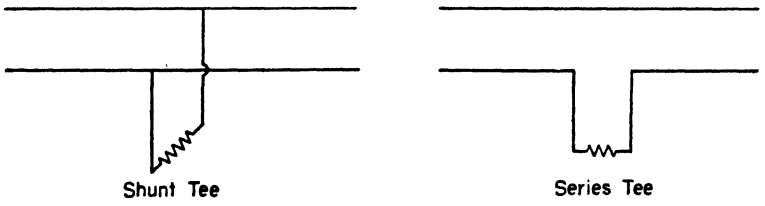


Fig. 7.16. Approximate equivalent circuits of waveguide tee-sections.

(1) Series T or E type.
(2) Shunt T or H type.

The approximate equivalent circuits are shown in Fig. 7.16.

When the direction of the waveguide must be changed, circular bends are ordinarily used. A circular bend is one in which the internal dimensions of the waveguide are kept constant as its direction is changed by some angle ϕ. The two basic types of bends are shown in Fig. 7.17. An E-bend is so called because the electric field vector

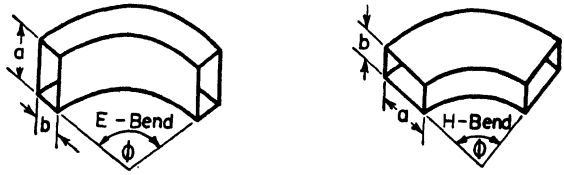


Fig. 7.17. Waveguide bends.

is rotated through an angle ϕ in passing through the bend. In the same manner, an H-bend is one in which the magnetic field intensity vector is rotated through ϕ degrees.

Instead of using a circular bend to change the direction of a waveguide, an abrupt change could be made. Of course, such a sharp change in direction would cause standing waves to be set up due to reflection. However, by inserting a plane directly in the corner,

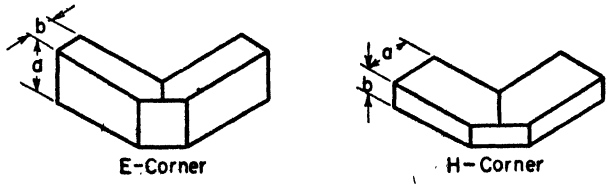


Fig. 7.18. Waveguide corners.

these reflections are eliminated. As in the case of bends, there are two types, E-corners and H-corners. They are shown in Fig. 7.18.

In some applications it becomes necessary to change the polarization of the field in the guide. This is readily accomplished by twisting the guide as indicated in Fig. 7.19. Generally, the twist should be rather gradual in order to minimize reflections.

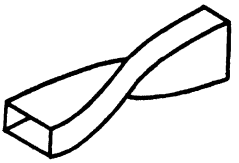


Fig. 7.19. Waveguide twist.

Practical, specific design data on these sections are usually available from manufacturers of waveguide components.

7.14 Directional Couplers

A waveguide device of considerable practical significance is the *directional coupler*. Its function is primarily to separate the incident and reflected waves in a waveguide from one another. For this reason it is sometimes characterized by the term *wave selector*.

A directional coupler is usually designated as a one-hole, or a two-hole coupler. Such terms, in addition to the very generally used "plumbing" to designate waveguide assemblies, lead one to suspect the early training of some of the more imaginative engineers operating in this field.

The one-hole coupler consists of a short closed section of rectangular waveguide. This section is coupled to the main waveguide through a single hole in the wide dimension of the guide. A probe is placed at the end near the oscillator while the other end is covered with a material which will absorb *RF* energy. This arrangement is

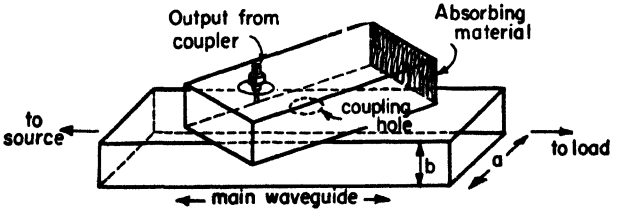


Fig. 7.20. One-hole directional coupler.

indicated in Fig. 7.20. The introduction of the hole in the main waveguide disturbs the usual electric field configuration at this point for two reasons:

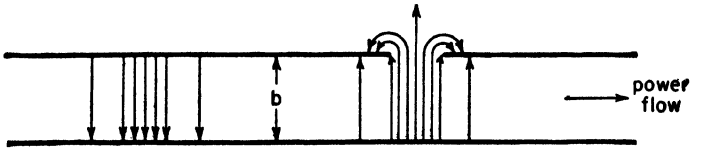


Fig. 7.21. Fringing of the electric field through a hole in the wide (a) dimension of the waveguide. The magnetic field is not shown.

(1) The electric field, which was originally normal to the hole, fringes as shown in Fig. 7.21.
(2) The surface currents produced by the magnetic field are interrupted by the hole so that they flow along the inner surface, through the hole, to the outer surface. This results in a second component of electric field intensity directed as shown in Fig. 7.22.

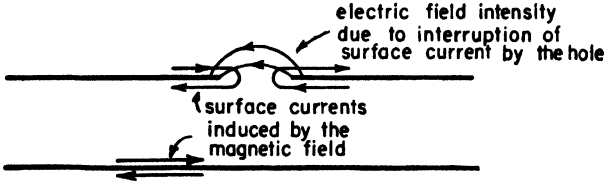


Fig. 7.22. Electric field produced by the surface currents.

These two electric field components on the outside of the hole reinforce each other on the side of the hole away from the direction of propagation. They tend to cancel one another on the side in the direction of propagation. This is indicated in Fig. 7.23. However,

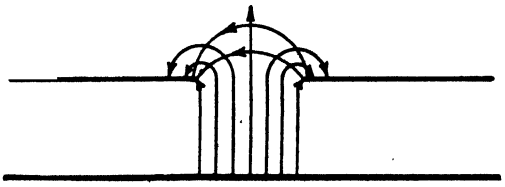


Fig. 7.23. Total electric field in the hole. Superposition of Figs. 7.21 and 7.22.

the magnetic field about the hole is symmetrical. As a result, from Poynting's theorem, it is evident that, of the total power which flows out of the main waveguide through the hole into the closed section,

more flows to the left (opposite to the direction of propagation in the main waveguide) than flows to the right (same as the direction of propagation). This argument is based upon the existence of the incident wave only, since propagation in only one direction was considered. Thus, by placing a pickup probe at the left end, and a layer of absorbing material at the right, the power may be measured.

The power in the reflected wave does not enter directly into this analysis because, from a consideration of the fields that it will establish in the hole, it is evident that most of the reflected wave energy which passes through the hole goes to the right end of the coupler and is absorbed by the coating at the end. By properly selecting the size of the hole and by skewing the coupler, it is possible to establish the magnitudes of the electric field intensity components previously mentioned so that they practically cancel on one side of the hole. In such a case, the probe receives practically nothing but energy from the incident wave, while the reflected wave is absorbed by the material at the other end. Thus, the directional coupler can be used to measure the power in the incident wave, or, by reversing the coupler, the power in the reflected wave.

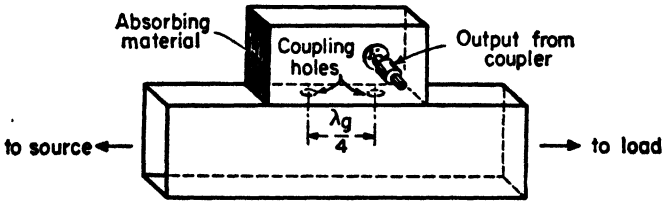


Fig. 7.24. Two-hole directional coupler.

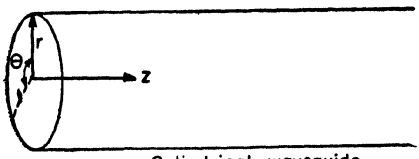
A typical two-hole coupler is shown in Fig. 7.24. Like the one-hole coupler, it consists of a closed section of waveguide with a probe at one end, and *RF* energy-absorbing material at the other. However, it is mounted on the narrow dimension of the main waveguide and coupled to it through two holes a quarter wavelength apart. Because of the location of the probe, relative to the direction of propagation, it is evident that the components of the incident wave that pass through the two holes and travel toward the probe arrive there in phase with one another since they travelled the same distance. Hence, upon arrival at the probe they add directly. On the other hand, the portions of the reflected wave that pass through

the holes cancel at the probe because one part had to travel a half wavelength further and had its phase shifted by 180° in the process. Hence, the reflected wave has no effect on the measuring device. Parts of the waves which do not go to the probe are removed by the absorbing material at the other end.

7.15 Cylindrical Waveguides

The discussion so far has been confined to rectangular waveguides for two reasons:

(1) They are currently of the greatest practical importance.
(2) Their analysis is easier to accomplish mathematically, and their operation is easier to visualize physically, than other types of waveguides. The concepts and techniques that have been developed are perfectly general and may be extended directly to the more mathematically difficult cylindrical waveguide.



Cylindrical waveguide

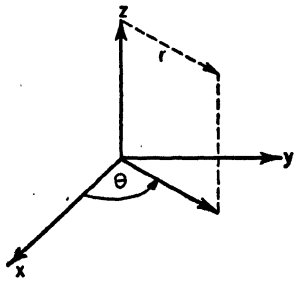


Fig. 7.25. Relationships between cylindrical and rectangular coordinate systems.

The chief obstacle to mathematical convenience in the analysis of the cylindrical waveguide is the necessity of using cylindrical coordinates. In order to apply the boundary conditions at a fixed value of one coordinate, namely r, cylindrical coordinates must be used. The relationship between rectangular and cylindrical coordinates is shown in Fig. 7.25. Hence,

$$x = r\cos\theta \tag{7.72a}$$

$$y = r\sin\theta \tag{7.72b}$$

$$z = z \tag{7.72c}$$

$$r = \sqrt{x^2 + y^2} \tag{7.72d}$$

The wave equation in rectangular coordinates was shown to be

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \gamma^2 + k^2\right)\mathbf{E} = 0$$

Assuming that **E**, the electric field intensity, is some function of r and θ, that is, of x and y, then let

$$E = E(r,\theta)$$

where E is understood to mean E_X, E_Y, or E_Z, one of the components of **E**. The derivative with respect to x is then

$$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial E}{\partial \theta}\frac{\partial \theta}{\partial x} \tag{7.73}$$

But, since

$$r = \sqrt{x^2 + y^2} \qquad \text{and} \qquad \theta = \tan^{-1}\left(\frac{y}{x}\right)$$

it then follows that

$$\frac{\partial r}{\partial x} = \frac{x}{r} \qquad \frac{\partial \theta}{\partial x} = -\frac{y}{r^2}$$
$$\frac{\partial r}{\partial y} = \frac{y}{r} \qquad \frac{\partial \theta}{\partial y} = \frac{x}{r^2}$$

Hence, the derivative of the electric field intensity with respect to x is

$$\frac{\partial E}{\partial x} = \frac{x}{r}\frac{\partial E}{\partial r} - \frac{y}{r^2}\frac{\partial E}{\partial \theta} \tag{7.74}$$

Now, taking the second derivative yields

$$\frac{\partial^2 E}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial E}{\partial r}\frac{x}{r}\right) - \frac{\partial}{\partial x}\left(\frac{\partial E}{\partial \theta}\frac{y}{r^2}\right) \tag{7.75}$$

Considering each term in this equation separately yields

$$\frac{\partial}{\partial x}\left(\frac{\partial E}{\partial r}\frac{x}{r}\right) = \left(\frac{\partial E}{\partial r}\right)\frac{\partial}{\partial x}\left(\frac{x}{r}\right) + \left(\frac{x}{r}\right)\frac{\partial}{\partial x}\left(\frac{\partial E}{\partial r}\right)$$
$$\frac{\partial}{\partial x}\left(\frac{\partial E}{\partial \theta}\frac{y}{r^2}\right) = \left(\frac{\partial E}{\partial \theta}\right)\frac{\partial}{\partial x}\left(\frac{y}{r^2}\right) + \left(\frac{y}{r^2}\right)\frac{\partial}{\partial x}\left(\frac{\partial E}{\partial \theta}\right)$$

The indicated derivatives in the expressions above are evaluated below.

$$\frac{\partial}{\partial x}\left(\frac{x}{r}\right) = \frac{\partial}{\partial x}\left(\frac{x}{\sqrt{x^2 + y^2}}\right) = \frac{y^2}{r^3}$$
$$\frac{\partial}{\partial x}\left(\frac{\partial E}{\partial r}\right) = \frac{\partial^2 E}{\partial r^2}\frac{\partial r}{\partial x} + \frac{\partial^2 E}{\partial r\,\partial \theta}\frac{\partial \theta}{\partial x} = \frac{x}{r}\frac{\partial^2 E}{\partial r^2} - \frac{y}{r^2}\frac{\partial^2 E}{\partial r\,\partial \theta}$$
$$\frac{\partial}{\partial x}\left(\frac{y}{r^2}\right) = \frac{\partial}{\partial x}\left(\frac{y}{x^2 + y^2}\right) = -\frac{2xy}{r^4}$$
$$\frac{\partial}{\partial x}\left(\frac{\partial E}{\partial \theta}\right) = \frac{\partial^2 E}{\partial \theta^2}\frac{\partial \theta}{\partial x} + \frac{\partial^2 E}{\partial \theta\,\partial r}\frac{\partial r}{\partial x} = -\frac{y}{r^2}\frac{\partial^2 E}{\partial \theta^2} + \frac{x}{r}\frac{\partial^2 E}{\partial r\,\delta\theta}$$

Substituting these relations back into Eq. (7.75) for the second derivative, and rearranging terms somewhat, yields

$$\frac{\partial^2 E}{\partial x^2} = \left(\frac{x^2}{r^2}\right)\frac{\partial^2 E}{\partial r^2} + \left(\frac{y^2}{r^3}\right)\frac{\partial E}{\partial r} + \left(\frac{y^2}{r^4}\right)\frac{\partial^2 E}{\partial \theta^2} + \left(\frac{2xy}{r^4}\right)\frac{\partial E}{\partial \theta} - \left(\frac{2xy}{r^3}\right)\frac{\partial^2 E}{\partial r\ \partial \theta} \tag{7.76}$$

Now, the same process may be followed for the second derivative with respect to y, obtaining

$$\frac{\partial^2 E}{\partial y^2} = \left(\frac{y^2}{r^2}\right)\frac{\partial^2 E}{\partial r^2} + \left(\frac{x^2}{r^3}\right)\frac{\partial E}{\partial r} + \left(\frac{x^2}{r^4}\right)\frac{\partial^2 E}{\partial \theta^2} - \left(\frac{2xy}{r^4}\right)\frac{\partial E}{\partial \theta} + \left(\frac{2xy}{r^3}\right)\frac{\partial^2 E}{\partial r\ \partial \theta} \tag{7.77}$$

If the two second derivatives are added together, it will be noted that the last two terms cancel so that

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} = \left(\frac{x^2 + y^2}{r^2}\right)\frac{\partial^2 E}{\partial r^2} + \left(\frac{x^2 + y^2}{r^3}\right)\frac{\partial E}{\partial r} + \left(\frac{x^2 + y^2}{r^4}\right)\frac{\partial^2 E}{\partial \theta^2} \tag{7.78}$$

But, since $x^2 + y^2 = r^2$, then

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} = \frac{\partial^2 E}{\partial r^2} + \frac{1}{r}\frac{\partial E}{\partial r} + \frac{1}{r^2}\frac{\partial^2 E}{\partial \theta^2} \tag{7.79}$$

and the wave equation in cylindrical coordinates is

$$\left[\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} + (\gamma^2 + k^2)\right]E = 0 \tag{7.80}$$

where E stands for any rectangular component of $\mathbf{E}$.*

Since this is a differential equation in the two variables r and θ, the best technique of solution is the method of separation of variables already outlined in Art. 7.2. Assume that the electric field intensity may be expressed implicitly as the product of two functions, one $R(r)$ of r only, and one $\Theta(\theta)$ of θ only. Hence,

$$E(r_1\theta) = R(r)\,\Theta(\theta) \tag{7.81}$$

or, omitting the arguments of the functions,

$$E = R\Theta \tag{7.82}$$

Substituting in the wave equation yields

$$\Theta\frac{\partial^2 R}{\partial r^2} + \frac{\Theta}{r}\frac{\partial R}{\partial r} + \frac{R}{r^2}\frac{\partial^2 \Theta}{\partial \theta^2} + (\gamma^2 + k^2)R\Theta = 0$$

* See Appendix II.

Divide through by $R\Theta$ and then multiply through by r^2. The result is

$$\left[\frac{r^2}{R}\frac{\partial^2 R}{\partial r^2} + \frac{r}{R}\frac{\partial R}{\partial r} + r^2(\gamma^2 + k^2)\right] + \left[\frac{1}{\Theta}\frac{\partial^2 \Theta}{\partial \theta^2}\right] = 0 \quad (7.83)$$

The equation now consists of two terms:

(1) A function of r only — the first bracket.
(2) A function of θ only — the second bracket.

Consequently, the variables are separated.

Following the procedure outlined in Art. 7.2 yields

$$\frac{r^2}{R}\frac{\partial^2 R}{\partial r^2} + \frac{r}{R}\frac{\partial R}{\partial r} + K_1^2 r^2 = p^2 \qquad \frac{1}{\Theta}\frac{\partial^2 \Theta}{\partial \theta^2} = K_2^2 \quad (7.84)$$

Therefore, the equation in r becomes

$$r^2\frac{\partial^2 R}{\partial r^2} + r\frac{\partial R}{\partial r} + (r^2K_1^2 - p^2)R = 0 \quad (7.85)$$

where $$K_1^2 = \gamma^2 + k^2$$

This is a form of Bessel's equation and has a solution of the form

$$R = C_1 J_p(rK_1) + C^2 Y_p(rK_1) \quad (7.86)$$

where $J_p(rK_1)$ and $Y_p(rK_1)$ are Bessel functions of order p, of the first and second kind. C_1 and C_2 are arbitrary constants.

The solution for Θ is readily obtained since it is an ordinary second-order linear differential equation with constant coefficients. Hence,

$$\Theta = C_3 \sin K_2\theta + C_4 \cos K_2\theta \quad (7.87)$$

Thus, the four possible solutions are

$$E_1 = A_1 J_p(rK_1) \sin K_2\theta \quad (7.88a)$$
$$E_2 = A_2 J_p(rK_1) \cos K_2\theta \quad (7.88b)$$
$$E_3 = A_3 Y_p(rK_1) \sin K_2\theta \quad (7.88c)$$
$$E_4 = A_4 Y_p(rK_1) \cos K_2\theta \quad (7.88d)$$

The proper solution and the values of the constants K_1 and K_2, as well as p, are determined from the boundary conditions. The arbitrary constant A is established by the generator.

The author feels that, from this point on, the techniques and results obtained are so closely analogous to the case of the rectangular

waveguide that their inclusion here would be unduly repetitious. The results may be found in any one of several handbooks, or in the references listed at the end of the chapter. Really, the only distinction between the rectangular and cylindrical cases is the occurrence of the Bessel function in the latter instance. The trepidation with which most engineers approach Bessel functions is actually not warranted, being due largely to lack of familiarity.*

7.16 Introduction to Cavity Resonators

Just as the waveguide is the microwave counterpart of the transmission line, it is reasonable to expect the resonant line to have an analogous waveguide device. The corresponding element is the cavity resonator which, in its simplest form, consists of a short section of waveguide, short-circuited at one or both ends.

At microwave frequencies, the constants of an ordinary tuned circuit are so small that they are not physically realizable in lumped form. Furthermore, the frequency is occasionally so high that the resonant transmission line may also have impractically small dimensions. Consequently, although resonant coaxial lines are in common use at microwave frequencies, conditions arise which may require the application of resonant sections of waveguide.

A tuned circuit is ordinarily characterized in terms of three unique quantities:

(1) The Q.
(2) The resonant frequency.
(3) The shunt resistance at resonance R_{sh}.

Unfortunately, a resonant cavity cannot be uniquely specified in terms of these quantities because

(1) It can oscillate in an infinite number of modes.
(2) There are no unique values of inductance and capacitance.

Nevertheless, certain restrictions may be made under which the Q, resonant frequency, and shunt resistance at resonance for cavities are significant quantities. These will be discussed separately in the succeeding articles.

* Readers interested in Bessel functions are referred to *Bessel Functions for Engineers*, by McLachlan, Oxford Press, 1934, or *Mathematical Methods in Engineering*, by Karman and Biot, McGraw-Hill, 1940.

7.17 Technique of Solution, Rectangular Cavity

Consider the closed, rectangular cavity shown in Fig. 7.26 which is assumed to be made of metal sheets having infinite conductivity. Then, assuming the usual set of coordinate axes, for an electromagnetic wave to exist in the cavity, it must satisfy the wave equation, which is

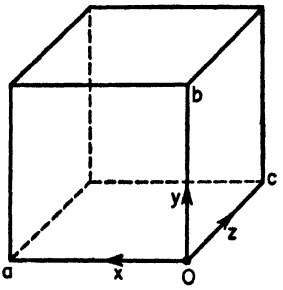


Fig. 7.26. Coordinates of the rectangular cavity.

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} + k^2\right)\mathbf{E} = 0$$

The boundary conditions that must be met are

$$\begin{aligned}
E_x &= 0 \quad \text{when } y = 0 \quad \text{and when } z = 0 \\
E_x &= 0 \quad \text{when } y = b \quad \text{and when } z = c \\
E_y &= 0 \quad \text{when } x = 0 \quad \text{and when } z = 0 \\
E_y &= 0 \quad \text{when } x = a \quad \text{and when } z = c \\
E_z &= 0 \quad \text{when } x = 0 \quad \text{and when } y = 0 \\
E_z &= 0 \quad \text{when } x = a \quad \text{and when } y = b
\end{aligned}$$

Using the method of separation of variables, let

$$E = X(x)\, Y(y)\, Z(z) \tag{7.89}$$

where E stands for any component of $\mathbf{E}$. Consequently, the wave equation becomes

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} + \frac{1}{Z}\frac{d^2Z}{dz^2} + k^2 = 0$$

and therefore,

$$\frac{1}{X}\frac{d^2X}{dx^2} = K_1^2 \qquad \frac{1}{Y}\frac{d^2Y}{dy^2} = K_2^2 \qquad \frac{1}{Z}\frac{d^2Z}{dz^2} = K_3^2 \tag{7.90}$$

The following solutions are possible:

$$X = A_1 \sin K_1 x + A_2 \cos K_1 x \tag{7.91a}$$
$$Y = B_1 \sin K_2 y + B_2 \cos K_2 y \tag{7.91b}$$
$$Z = C_1 \sin K_3 z + C_2 \cos K_3 z \tag{7.91c}$$

From inspection of the boundary conditions it is evident that

$$E = XYZ = D_1 \sin K_1 x \sin K_2 y \sin K_3 z \tag{7.92}$$

and that

$$K_1 = \frac{l\pi}{a} \tag{7.93}$$

$$K_2 = \frac{m\pi}{b} \tag{7.94}$$

$$K_3 = \frac{n\pi}{c} \tag{7.95}$$

Thus, any component of the electric field intensity is given by

$$E = D_1 \sin\left(\frac{l\pi}{a}\right)x \sin\left(\frac{m\pi}{b}\right)y \sin\left(\frac{n\pi}{c}\right)z \tag{7.96}$$

By the same method, any component of the magnetic field intensity is given by

$$H = D_2 \cos\left(\frac{l\pi}{a}\right)x \cos\left(\frac{m\pi}{b}\right)y \cos\left(\frac{n\pi}{c}\right)z \tag{7.97}$$

Substituting Eqs. (7.90) into the wave equation yields

$$K_1^2 + K_2^2 + K_3^2 + k^2 = 0$$

or, rearranging terms and extracting the square root,

$$k = j\sqrt{K_1^2 + K_2^2 + K_3^2} = j\beta$$

Then, substituting for the constants under the radical yields

$$\beta = \sqrt{\left(\frac{l\pi}{a}\right)^2 + \left(\frac{m\pi}{b}\right)^2 + \left(\frac{n\pi}{c}\right)^2}$$

But $$\beta = \frac{2\pi}{\lambda}$$

So $$\lambda = \frac{2\pi}{\beta} = \frac{2}{\sqrt{(l/a)^2 + (m/b)^2 + (n/c)^2}} \tag{7.98}$$

Thus, one resonant wavelength will be found when l, m, and n are integers, and there will be an infinite number of such points of resonance.

7.18 Resonant Frequency of a Cavity

The resonant frequency of a cavity is obtained from the solution of the wave equation (see Art. 7.17) and the application of the known boundary conditions to that solution. An infinite number of solutions will be obtained, as occurred in the case of the waveguide, one solution corresponding to each particular configuration of electric and magnetic fields. For each such mode there is one resonant frequency. This resonant frequency is dependent only upon the dimensions of the cavity. According to Eq. (7.98) for a rectangular cavity,

$$\lambda = \frac{2}{\sqrt{(l/a)^2 + (m/b)^2 + (n/c)^2}}$$

However,

$$f = \frac{v_1}{\lambda} = \frac{1}{\lambda\sqrt{\mu\epsilon}}$$

Therefore, the resonant frequency is

$$f = \sqrt{\frac{(l/a)^2 + (m/b)^2 + (n/c)^2}{\mu\epsilon}} \tag{7.89}$$

This solution is applicable only to the rectangular cavity. As a matter of fact, an analytical solution is possible only for a few special cavity shapes, ordinarily those which can be easily defined in terms of a standard system of coordinates. Unfortunately, these "regular" cavities do not always give the desired operating characteristics, and irregular forms are used. Since exact analytical solutions are next to impossible, a number of approximate methods of calculation have been developed that give reasonable engineering accuracy. Such data may be found in standard handbooks on the subject.

7.19 Cavity Q

The Q of any resonant system depends upon the energy stored in the system and the power losses. It can be uniquely defined as

$$Q = \left(\frac{\text{reactive volt-amperes in } L \text{ or } C}{\text{power lost}}\right) \text{ at resonance}$$

For a parallel R-L-C circuit, substitution into this equation for Q yields

$$Q = \frac{(E^2/\omega L)}{(E^2/R)}$$

It may be rewritten in a slightly different form as

$$Q = \omega \frac{(E^2/\omega^2 L)}{(E^2/R)}$$

The numerator and denominator are now in a form such that

$$\frac{E^2}{\omega^2 L} = LI^2 = \frac{1}{2} L(\sqrt{2}\, I)^2 = \text{maximum energy stored in } L$$

$$\frac{E^2}{R} = \text{power loss in } R$$

Hence, the equation for Q may be written as

$$Q = \omega\left(\frac{\text{maximum energy stored}}{\text{power loss}}\right)$$

A definition of this character is essential in the study of resonant cavities because there is no way of specifying particular values of inductance or capacitance, and that precludes the possibility of defining Q in terms of "circuit constants." Obviously, the definition of Q just given is unique only for one mode at a time, indicating that Q will vary with the modal configuration.

The power loss in a resonant cavity is due principally to conductor losses because most practical dielectrics have a negligible leakage conductance. The losses are I^2R losses where I is the circulating current in the cavity walls and R is the resistance of the current path. The current is assumed to be confined to the skin depth δ, calculated in Chap. 6, which had a value

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}}$$

where σ is the conductivity of the metal walls, μ is the permeability of the metal walls, and ω is the frequency in radians per second. The general expression for resistance is

$$R = \rho \frac{L}{A} = \frac{1}{\sigma} \frac{L}{A}$$

where ρ = resistivity = $1/\sigma$. Then the resistance of a differential length and cross section is

$$R = \frac{\rho}{\delta} \frac{dl}{ds} = \frac{1}{\sigma\delta} \frac{dl}{ds}$$

where dl is the length of the current path, ds is the width of current path, δ is the depth of current path, or skin depth, and $\delta\, ds$ is the area normal to current flow. The current flowing in the wall is equal to the magnetomotive force about the surface element carrying current. That is,

$$I = H\, ds$$

where H is the rms value of magnetic field intensity along the surface formed by ds and dl. Consequently, the power loss is

$$P = I^2R = H^2(ds)^2 \frac{1}{\sigma\delta} \frac{dl}{ds}$$

or

$$P = \frac{1}{\sigma\delta} H^2\, da$$

where $da = ds\, dl$ = differential *surface* area (not cross-section). Therefore, the total power loss would be

$$P = \int_s \left(\frac{1}{\sigma\delta}\right) H^2\, da$$

The maximum energy stored in a volume dV may be calculated from the basic equation for energy in the magnetic field, that is,

$$dW = \mu H^2\, dV$$

where H is the rms value of magnetic field intensity. Then, integrating, the total energy stored within the cavity volume may be obtained as

$$W = \int_V \mu H^2\, dV$$

Substituting these expressions for energy stored and power lost into the equation for the Q yields

$$Q = \omega \frac{W}{P} = \frac{(\omega\mu\sigma)\,\delta \int_V H^2\, dV}{\int_s H^2\, da}$$

If the magnetic field were constant throughout the cavity volume, then H^2 can be factored out of each integral and cancelled, giving

$$Q = \left(\frac{2}{\delta}\right)\left(\frac{\text{cavity volume}}{\text{surface area}}\right) \qquad \text{since } \omega\mu\sigma = \frac{2}{\delta^2}$$

As the result of the approximations made, this equation will always give a value for Q higher than that actually obtainable. It is a useful formula, nevertheless, for making a few general deductions about cavities.

Since the skin depth δ varies inversely as the square root of frequency, then the cavity Q will vary directly with the square root of the resonant frequency. Another general deduction may be drawn from the equation because of the presence of the factor (volume/area). It is apparent that large cavities will tend to have higher Q's than smaller ones.

It is evident that a cavity resonator differs in many respects from the conventional tuned circuit. For the latter, a high Q ordinarily implies a high shunt resistance at the resonant frequency. This may or may not be true for the resonant cavity, because it can be shown that the shunt impedance is inversely proportional to the square root of the resonant frequency, whereas the Q is directly proportional to the same quantity. Hence, it is possible to find cavities with almost any combination of shunt resistance and Q.

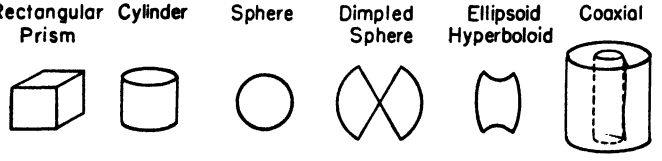


Fig. 7.27. Types of cavity resonators (adapted from Table X-1, *Microwave Transmission Design Data*, Sperry Gyroscope Co.).

7.20 Resonator Types

See Fig. 7.27 for a table of prototype resonator configurations.

PROBLEMS

7.1 A rectangular waveguide for which $a = 7$ cm, $b = 3.5$ cm, $\lambda_a = 10$ cm, the mode is $TE_{1,0}$, and attenuation is negligible, is supplied by a generator which sets up a magnetic field having a maximum value of 0.5 amp/cm in the center of the waveguide. Calculate (a) the average power transmitted, (b) the peak voltage across the y dimension of the waveguide.

7.2 A rectangular waveguide for which

$$a = 7 \text{ cm} \qquad b = 1 \text{ cm} \qquad \lambda_a = 10 \text{ cm} \qquad (TE_{1,0} \text{ mode})$$

is to be joined to another waveguide for which $a = 7.5$ cm. What must be

the b dimension of the second waveguide if there is to be no reflection from the junction?

7.3 An oscillator supplies power to a rectangular waveguide for which

$$a = 7 \text{ cm} \qquad b = 1 \text{ cm} \qquad \lambda_a = 10 \text{ cm} \qquad (\text{TE}_{1,0} \text{ mode})$$

The waveguide feeds an antenna whose input impedance is $50 + j23$ ohms. An inductive window is to be placed near the antenna and adjusted to remove standing waves. What is the minimum distance from the antenna, and what susceptance should the window introduce, to eliminate standing waves?

7.4 An E-type of T-junction is located 4 wavelengths behind the window in the waveguide assembly of Problem 7.3. Although both legs of the T-junction are terminated in impedances equal to the characteristic impedance of the waveguide, standing waves are found to exist between the oscillator and the T-junction.

(a) Explain the existence of the standing waves.

(b) Find the location and susceptance of an inductive window which will eliminate the standing wave.

REFERENCES

Books

Brainerd, J. G., Koehler, G., Reich, H. J., and Woodruff, I., *Ultra-High-Frequency Techniques*, Van Nostrand, 1942.

Reference Data For Radio Engineers, Federal Telephone and Radio Corp., 1946.

M.I.T. Staff, *Principles of Radar*, McGraw-Hill, 1946.

Sarbacher, R. I., and Edson, W. A., *Hyper and Ultrahigh Engineering*, Wiley, 1943.

Skilling, H. H., *Fundamentals of Electric Waves*, Wiley, 1942.

Slater, J. C., and Frank, N. H., *Electromagnetism*, McGraw-Hill, 1947.

Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill, 1941.

Microwave Transmission Design Data, Sperry Gyroscope Co., (23-80).

Periodicals

Barrow, W. L., "Transmission of Electromagnetic Waves in Hollow Pipes of Metal, *Proc. IRE*, **24**, 1298 (1936).

Brillouin, L., "Hyperfrequency Waves and Their Practical Use," *Elec. Comm.*, **19**, 18 (1941).

Carson, J. R., Mead, S. P., and Schelkunoff, S. A., "Hyperfrequency Waveguides—Mathematical Theory," *Bell System Tech. J.*, **15**, 2 (1936).

Chu, L. J., "Electromagnetic Waves in Hollow Elliptic Pipes of Metal," *J. App. Phys.*, **9**, 9 (1938).

Chu, L. J., and Barrow, W. L., "Electromagnetic Waves in Hollow Metal Pipes of Rectangular Cross Section," *Proc. IRE*, **26**, 1520 (1938).

Clavier, A. G., and Altovsky, V., "Experimental Researches on the Propagation of Electromagnetic Waves in Dielectric Guides," *Elec. Comm.*, **18**, 81 (1939).

Hansen, W. W., "A Type of Electrical Resonator," *J. Applied Phys.*, **9**, 654 (1938).

Linder, E. G., "Attenuation of Electromagnetic Fields in Pipes Smaller Than Critical Size," *Proc. IRE*, **30**, 554 (1942).

Southworth, G. C., "Hyper Frequency Waveguides—General Considerations and Experimental Results," *Bell System Tech. J.*, **15**, 284 (1936).

"Design Characteristics of Resonant Cavities," Westinghouse Research Report SR-127.

CHAPTER 8

UHF TRIODES AND OSCILLATORS

For many years the principal obstacle to the development of practical ultrahigh frequency systems of any type was the difficulty in generating the required ultrahigh frequency signals. Many advances have been made in recent years and all indications point toward further progress. This chapter discusses the principal factors in ordinary tubes that tend to prevent the production of ultrahigh frequency oscillations. Methods of minimizing these effects and refining the tube construction are then mentioned. This is followed by an analysis of practical oscillators suitable for use in the ultrahigh frequencies. The discussion is generally limited to conventional negative grid oscillator circuits, but since improvement in positive grid oscillators may occur, a short introductory article is included on that subject.

8.1 Triode High-frequency Equivalent Circuit

In the ultrahigh frequencies, where the wavelengths are very short, the electrode structure inside the glass envelope enclosing the tube is complicated by the inevitable presence of distributed capacitance and inductance. The capacitances between electrodes and the inductances of the leads from the pin connections to the electrodes themselves can no longer be considered negligible. Any accurate representation of the tube must include these parameters.

Fortunately, and to a reasonably high degree of accuracy, the distributed constants can be represented by lumped elements because, even at the highest frequencies, the lead lengths seldom exceed a quarter wavelength, and the longest electrode dimension is rarely more than a small fraction of a wavelength. Therefore, it is possible to represent a triode at the ultrahigh frequencies by the circuit shown in Fig. 8.1 where the triode is considered to have lumped capacitance between each pair of tube electrodes and lumped inductance in series with each electrode lead.

It would be entirely possible to develop an equivalent plate circuit theorem for a tube in which the distributed elements indicated in Fig. 8.1 are considered. However, the resulting equations are comparatively so complicated that such a development tends to confuse rather than to clear up the situation. A brief review of the physical nature of the problems posed is more to the point and ordinarily more instructive.

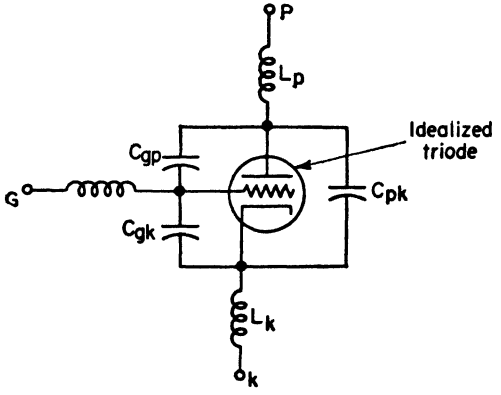


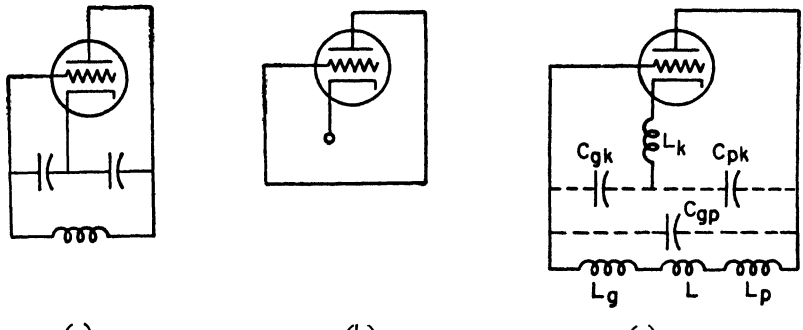
Fig. 8.1. Ultrahigh frequency equivalent triode circuit.

8.2 Frequency Limit of Ordinary Oscillators

The frequency of an oscillator is determined primarily by the constants in the tuned circuit and has a value

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

when the effects of losses in the oscillating circuit are neglected. A typical oscillator circuit is shown in Fig. 8.2(a) in which only essential *RF* components are indicated. In order to increase the frequency of oscillation, the magnitudes of the inductances and capacitances in the tank circuit must be reduced. This reduction can be carried to the point indicated in Fig. 8.2(b) where the physical lumped capacitances have been replaced by apparently open circuits, and the inductance by an ordinary metallic conductor. The diagram hardly appears as a complete circuit. However, from Art. 8.1 it is known that interelectrode capacitances are present and the grid-to-plate



(a) (b) (c)

Fig. 8.2. (a) Essential components of a typical oscillator. (b) Oscillator in which the tank circuit L and C have been reduced to the greatest extent. (c) Actual oscillator circuit of (b) when distributed tube elements are added.

conductor has a lead inductance associated with it. Consequently, the actual circuit appears as shown in Fig. 8.2(c).

In the limit, the minimum values of L and C in the oscillating circuit are the inductance that is inherent in the plate, grid, and cathode leads, and the capacitance that exists between tube leads and electrodes. Thus, the extent to which the constants of the oscillating circuit can be reduced imposes one restriction upon the upper operating frequency limit and is ordinarily determined by the geometry of the tube and associated circuit.

The method of raising the frequency of operation by changing the constants in the tuned circuit produced upper limits between 10 and 60 mcps for ordinary tubes, but reached as high as 700 mcps with special tubes such as the WE 316A. Other specially constructed tubes, involving advanced manufacturing techniques, have pushed this limit up to about 3000 mcps.

A second limitation upon the maximum frequency of oscillation of conventional oscillator circuits is imposed by the reduction in tank circuit Q at elevated frequencies. This lowering of the circuit Q as a function of frequency is due principally to increased

(1) Skin effect.
(2) Dielectric losses in the tube base and envelope.
(3) Radiation from the oscillating circuit.

A third restriction upon the upper frequency of oscillation arises due to electron transit time. When the transit time of the electron from cathode to plate is a certain fraction of the *RF* period, the grid

develops an in-phase component of current requiring that the grid supply power. This grid loading, which will be discussed at greater length in Art. 8.5, reduces the oscillator power output to impractically small values at frequencies in the neighborhood of 40 or 50 mcps, assuming operation with conventional tubes.

The reduction in tank circuit Q could be counterbalanced by increasing the transconductance of the tube. This could be accomplished by raising the operating plate voltage, which, in addition, would tend to reduce transit time. Unfortunately, the increased plate dissipation could be accomodated only by increasing the area of the plate. This tends to increase the interelectrode capacitance and in turn would partially nullify the advantageous effects of the increased plate voltage.

8.3 Effects of Interelectrode Capacitances

The most immediately obvious effect of excessive interelectrode capacitances is the lowering of the maximum possible frequency at which the tube can oscillate. Reduction of this capacitance will naturally allow the generation of higher frequencies.

There are other undesirable effects resulting from the presence of unduly large tube capacitances. Generally, their reactances are so low at the frequencies of desired operation that they cause excessive circulating currents which increase the I^2R losses in the tank circuit. These currents form practically all of the circulating current in the tuned circuit and are much larger than the power currents delivered by the tube.

Because of the character of these capacitances, they are not constant, but vary in a complex manner with the oscillator loading and the electrode voltages. Consequently, for frequency stability, they should not have too much control over the operating frequency of the oscillator.

Moreover, the grid-to-plate capacitance C_{gp} may, through its coupling of grid circuit to plate circuit, cause the input conductance of the tube to be undesirably large. Furthermore, this conductance may be either positive or negative, depending principally upon the impedance in the plate circuit. Proof of this phenomenon is omitted since it may be found in any one of many standard books on vacuum tube fundamentals.

8.4 Effects of Lead Inductance

Due to the presence of lead inductances, it is impossible to make a direct connection to any of the tube electrodes. Consequently, the lead inductance forms a part of the oscillator tank circuit which, at the ultrahigh frequencies, is a large proportion of the total inductance present. It is not particularly objectionable except that it imposes an upper frequency limit on the tube as an oscillator.

It can be shown by a very tedious mathematical analysis of the equivalent tube circuit that the lead inductance, particularly the cathode lead inductance, has the effect of increasing the input and output conductances of the tube. The input conductance is generally more important than the output conductance, the latter frequently being negligible in comparison. A thorough study of the problem indicates that the upper frequency limit for screen grid and other multielectrode tubes is set by the lead inductance, whereas it is determined by transit time in triodes.

As an approximation, it can be shown that the input conductance is given by the following equation:

$$G_{in} = \omega^2 g_m L_K C_{g_K}$$

This relationship is very nearly true for pentodes, but is quite approximate for triodes. It is evident that since the input conductance varies as the square of the frequency, even slight extensions in operating frequency cause a marked increase in the input conductance.

8.5 Transit Time

The transit time of the electrons in their flight from cathode to plate becomes objectionable when this time becomes so large, compared to the period of the *RF* grid voltage, that irregularities in plate current cause different numbers of electrons to be approaching and receding from the grid at any given instant of time. This inequality in electrons on either side of the grid generates an induced grid current which causes the input admittance of the tube to have a real term. According to W. R. Ferris (see References), this conductance component of the input admittance varies as the square of the impressed frequency. At extremely high frequencies, where the electron transit time reaches a half or even a quarter of a period, this law no longer applies. The mathematical development of the effects of transit time would cover many pages of rather awesome appearing work.

For our purposes, a brief examination of the physical nature of the problem is perhaps more instructive.

The simplest way of stating the nature of the phenomenon is as follows. Assuming an ultrahigh frequency grid signal, then when the grid is more negative than normal, the number of electrons approaching the grid from the cathode is less than the number leaving the grid and progressing toward the plate. Such an unbalance of forces causes the grid to do work. On the other hand, during the next half cycle when the grid is more positive than normal, the number of electrons approaching the grid from the cathode is greater than the number in the grid-to-plate volume. Again, actual work is required of the grid. The necessity of delivering energy, thus imposed upon the grid, is met by the development of a component of grid current in phase with the applied grid voltage. As a result, a fairly large input conductance may exist due to the transit time effects. The work done by the grid is largely in accelerating the electrons and the power input to the grid will, therefore, appear as heat on the plate of the tube.

The difference between the number of electrons in the cathode-grid space and those in the grid-plate space can be kept small only by reducing the transit time of the electrons to a point where it is small compared to the period of the variations in plate current. This can be accomplished either by reducing the distance the electrons have to travel, or by increasing the velocity with which they make their trip. Thus, either the electrode spacing can be reduced, or the plate voltage can be increased, or both techniques may be used. There are limits, however, beyond which such modifications cannot be pushed. For example, the decrease in electrode spacing makes tube manufacture more difficult and, in order to keep the interelectrode capacitance from becoming extraordinarily large, the decrease in spacing must be accomplished by a decrease in electrode surface area. This, of course, reduces the allowable plate dissipation. The increase in plate voltage is limited by the reduction in transconductance which necessarily accompanies a decrease in plate current. The plate current has to be decreased, when the plate voltage is increased, in order to keep within the allowable plate dissipation.

In pulsed operation for use in radar, for example, transit time is generally reduced by using very high plate voltages during the

pulse. This is possible without plate current reduction because the plate dissipation during the pulse can be thousands of times greater than the average allowable dissipation due to the inherent thermal inertia of the plate structure. Thus, tubes may be run at higher frequencies when pulsed than when operated CW.

For pulsed operation, plate voltages of the order of ten thousand volts are used in order to obtain high peak powers and to reduce transit time. These voltages exert considerable influence upon the mechanical features of the tubes designs. The glass seals between electrodes are separated as much as possible in order to reduce the voltage gradient in the glass and on its surface. Also, sharp corners on high-voltage conductors are avoided to minimize the possibility of corona.

8.6 Minimizing Tube Capacitances and Inductances

The lead inductance and capacitance can be reduced by using shorter lengths of wire for the leads. This also reduces the circuit resistance. The requirement for small interelectrode capacitances in order to obtain high frequencies and low charging-current loss can be met by using small electrodes. The capacitance could also be decreased by increasing the spacing between the elements but, unfortunately, this increases the transit time.

Losses due to skin effect can be reduced by using low-resistance, large-surfaced conductors. Close spacing of the electrodes tends to reduce the radiation losses, but unfortunately, very close spacing increases the *RF* resistance. Dielectric losses are minimized by bringing the electrode leads directly through the glass envelope, doing away with the ordinary tube socket. This construction also reduces the lead inductances and capacitances since the leads are shorter.

Doubled-ended construction is also used frequently in UHF tube construction. In such tubes, each electrode has two separate terminals brought out individually through the glass envelope. These terminals are then connected symmetrically to the external circuit so that the lead inductances are effectively paralleled. Another technique is to arrange the leads inside the tube so that they are extensions of external resonant transmission lines. The tube elements are then maintained at the center of a resonant half-wave transmission line system with shorted ends. The lumped capacitance between electrodes is then assumed to be equally divided between the two quarter-wave sections on either side of the tube elements. Such tubes

will oscillate at higher frequencies because the capacitance associated with a single quarter-wave section has been materially reduced. The double-ended construction also reduces the radiation losses and the I^2R losses in the leads.

Grid lead inductance may be reduced by using parallel bars instead of the more common helically wound wire in the construction of the grid. In recent years, improved glass-working techniques have permitted the construction of tubes in which the electrode leads have been replaced by sealed metal disks. This makes the whole perimeter of the electrode available and allows the electrodes to be built into tuned circuits directly, thus yielding higher Q's and frequencies of oscillation. The *lighthouse tube* is typical of this *disk seal construction.*

There are other methods of reducing these various effects which make slight dissimilarities between the various UHF tubes. Without question, other techniques are being developed. The purpose of this article has been to discuss a few of the outstanding constructional features which make UHF tubes appear to be so different, externally, from ordinary tubes.

8.7 Plate Cooling

High-power operation of UHF triodes is hindered by the reduction in electrode area necessary for small interelectrode capacitances. This, and the reduction of transit time through the use of large plate voltages, makes the problem of heat dissipation from the plate quite difficult. The heat may be removed more effectively by adding cooling fins to the plate structure. This increases the thermal radiating surface without changing the tube capacitances. An air blast played over the plate will assist materially. For really high-power operation, water cooling is frequently used. Other tubes use plates made of tantalum and operate normally at red heat. This allows higher plate dissipations due to more rapid radiation of heat through operation at a higher temperature.

Another technique which is quite effective and rather common in current tubes is to use a solid block as the plate, bring it out externally to the tube proper by disk seal construction, and then provide it with thermal radiating fins. The small working surface area of the plate near the grid and cathode reduces the tube capacitances, but the large over-all area, plus the external exposure to air, increases the heat-dissipating ability of the tube.

8.8 Filament Emission

Very large filament emissions are generally required of UHF tubes for two reasons. First, the large *RF* currents in the plate-to-cathode and grid-to-cathode capacitances are, to a large extent, supplied by the cathode. Secondly, in pulsed operation, extremely large currents are required over short periods of time. Although this is a conduction current, and the first is a displacement current, both require actual cathode emission. Consequently, large filaments operating at high temperature are required, generally adding to the difficulty of tube cooling. Although oxide-coated cathodes are occasionally used in lower-power tubes, thoriated tungsten is most generally employed since it is less susceptible to positive ion damage and can safely provide the required degree of emission.

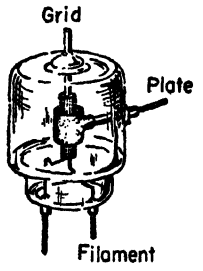
The use of thorium on the filaments does cause some difficulty, however, due to grid contamination resulting from evaporated thorium depositing on the grid. Since the grid is generally quite hot, actual grid emission may occur causing a d-c current to exist in the grid circuit.

8.9 Special UHF Tubes

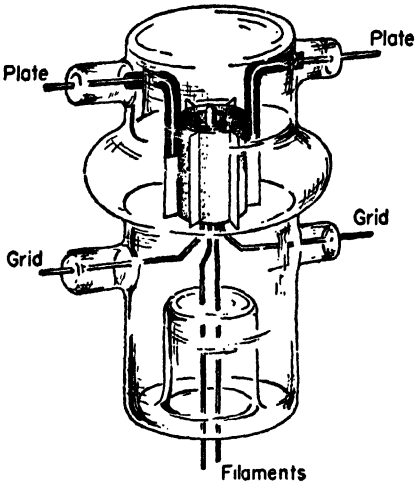
Because of the various factors outlined in the preceding articles, special tubes have been developed for various uses in the ultrahigh frequencies. The list given below is by no means complete, but it does give representative types.

Name	Type	Construction	Typical Service
GL 446 Lighthouse	triode	disk seal	Local osc. RF amplifier
GL 464 Lighthouse	triode	disk seal	UHF power osc.
WE 316A Doorknob	triode	single-ended	RF osc.
WE 368A Doorknob	triode	double-ended	RF osc.
708A Grounded grid	triode	single-ended	Mixer
811 Micropup	triode	single-ended grid, disk seal plate	RF osc.
832 and 829B	push-pull beam tetrode	single-ended	RF osc.
888	triode	single-ended water-cooled	RF osc.
954 Acorn	pentode	single-ended	RF amplifier
955 Acorn	triode	single-ended	Mixer
956 Acorn	pentode	single-ended	RF amplifier
GL 8012A	triode	double-ended	RF osc. & ampl.

Other commercially available lighthouse tubes are the GL 2C39, 2C40, 2C43, 3C22, and 5648.



Single ended construction
Note parallel grid wire construction.



Double ended construction
Note cooling fins on the plate.

Fig. 8.3. UHF triodes.

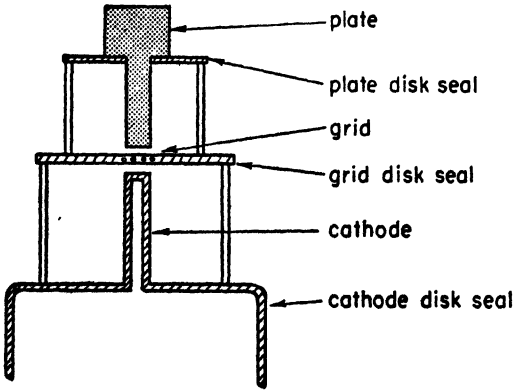


Fig. 8.4. Cross section (principal components only) of lighthouse triode.

Sketches of representative samples of these tubes appear in Figs. 8.3 and 8.4. Particular points to observe are the ways in which the

lead inductances and interelectrode capacitances have been reduced. Of particular interest is the lighthouse tube which is designed to fit into a coaxial transmission line tuner as shown schematically in Fig. 8.5.

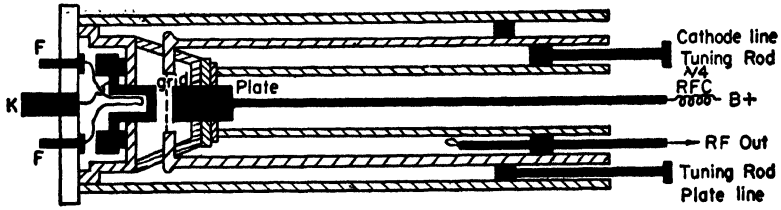


Fig. 8.5. Lighthouse tube and coaxial tuner.

8.10 Basic Oscillator Theory

The fundamental theory of feedback in general, and oscillators in particular, is ordinarily presented in considerable detail in the introductory courses on vacuum tubes. A very brief review of the subject will be introduced here to facilitate discussion of UHF circuits performing oscillator service.

The basic requirement that any oscillator circuit must fulfill is to completely regenerate a disturbance in its input, both in magnitude and phase. This can be done in either of two general ways:

(1) By external feedback connections on an amplifier.
(2) By developing a negative resistance characteristic internal to the device.

Circuits of the latter type were discussed in the introductory articles of Chap. 2. The material in this chapter is concerned solely with the first type, *feedback oscillators.*

The essential functional components of an oscillator may be listed for the sake of convenience as

(1) A regenerating circuit - to regenerate the original input.
(2) A frequency controller - to stabilize the frequency of oscillation to a particular desired frequency.
(3) A limiter - to maintain the amplitude of oscillation constant.

In a feedback oscillator, the regenerating circuit consists of an amplifier and an external feedback circuit. The frequency controller may be a resonant electrical circuit (*LC* tank circuit), a piezoelectric

circuit, a magnetostriction device, or a frequency-selective *R-C* circuit. UHF oscillators ordinarily use tuned electrical circuits. Moreover, for feedback oscillators in general, the functions of frequency control and feedback are customarily fulfilled by the same circuit.

The limiter may be the vacuum tubes in the amplifier, or it can be any other device or circuit having a nonlinear current-voltage characteristic. In this chapter it is assumed that all of the oscillators are limited by nonlinearities in the plate characteristics of the tubes in the amplifier section of the circuit.

Typical and representative feedback oscillators are the Colpitts, Hartley, Meissner, tuned plate, tuned grid, and Wien bridge. Most UHF oscillators are Colpitts or Hartley, with the Colpitts predominating.

The principle analytical method associated with oscillator analysis and design is called the *Barkhausen criterion.* It is derived very easily by letting

E_0 = output voltage from the amplifier
E_{fb} = voltage fed back into the input
β = feedback factor = $\dfrac{E_{fb}}{E_0}$
K_T = total amplification of the system from input to output terminals

From the definition of amplification we can write

$$K_T = \frac{E_0}{E_{\text{in}}}$$

But, for an oscillator, E_{in}, the input voltage, is equal to the voltage fed back, E_{fb}. Hence,

$$K_T = \frac{E_0}{E_{fb}}$$

or

$$K_T = \frac{1}{\beta}$$

This last relation is the fundamental mathematical form of the Barkhausen criterion. It is a relationship that is apparent even from a casual consideration of an oscillator circuit, simply stating that the voltage multiplication in the amplifier must exactly counterbalance the voltage-dividing ratio of the feedback circuit.

The procedure for analytical circuit solution is then as follows:

(1) Determine $\beta = \dfrac{E_{fb}}{E_0}$ in terms of the circuit constants.
(2) Determine K_T in terms of the circuit constants.
(3) Apply the Barkhausen criterion, that is, set $K_T = 1/\beta$.
(4) Equate reals and imaginaries on either side of this relation, obtaining two equations.
(5) Thus, since there are only two equations in any oscillator, all of the circuit parameters, except two, must be specified one way or another before the Barkhausen criterion is applied.

It is recognized that analytical methods are restricted to linear circuits for which the equivalent plate circuit theorem is applicable, and that most oscillators are inherently nonlinear in operation. However, there is much to be said for the Barkhausen criterion as an aid to a physical understanding of oscillator circuits.

8.11 Tapped Resonant Circuits

Most feedback oscillators consist of a single vacuum tube for amplification purposes and a tapped resonant circuit for feedback and frequency control. Typical tapped resonant circuits are shown in Fig. 8.6.

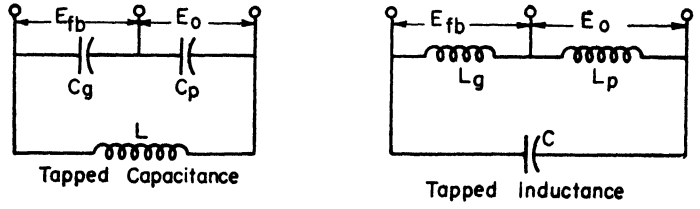


Fig. 8.6. Tapped resonant circuits.

Recognizing the fact that, under ideal conditions, there is an inherent 180° phase shift in the voltage as amplification occurs, then it is evident that for regeneration to occur, the feedback circuit must provide an additional 180° phase shift.

The manner in which the tapped resonant circuit provides this additional 180° phase shift may be explained by redrawing the circuits as shown in Fig. 8.7 and then constructing the vector diagrams of Fig. 8.8. For the sake of discussion, consider the circuit containing

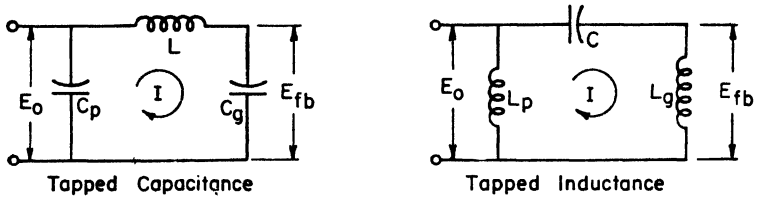


Fig. 8.7. Tapped resonant circuits redrawn.

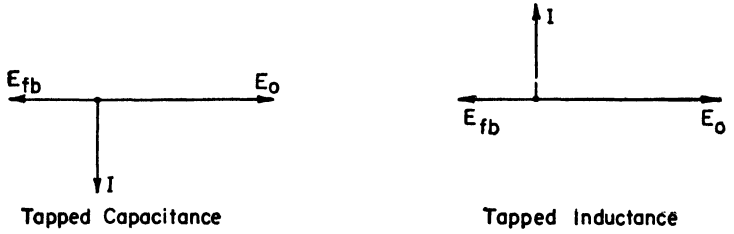


Fig. 8.8. Vector diagrams for the tapped resonant circuits of Fig. 8.7.

the tapped capacitance. Use the output voltage E_0 as the reference vector. The current through the leg containing the series combination of L and C_g will be lagging E_0 by 90° because, for resonance to occur, this arm must exhibit a net inductive reactance, equal in magnitude to the capacitive reactance of C_p. Then, the voltage across C_g lags this current by 90° as is characteristic of condensers. Thus, this voltage, E_{fb}, is 180° out of phase with E_0, the output voltage. The same analysis may be applied to the tapped inductance circuit to obtain the corresponding vector diagram.

Consequently, these circuits may be made effective feedback circuits for single-tube oscillators by connecting the cathode to the tap, and the grid and plate to the other two terminals.

8.12 Colpitts and Hartley Oscillators

Connection of the tapped resonant circuits to vacuum tubes as described in the preceding article yields the circuit diagrams of Fig. 8.9, which are the prototype Colpitts and Hartley oscillators. Only essential *RF* components are shown; all d-c paths have been omitted. Assuming that limiting is provided by overloading the tubes, it is apparent that these circuits contain all of the components listed in Art. 8.10 as essential to oscillator operation.

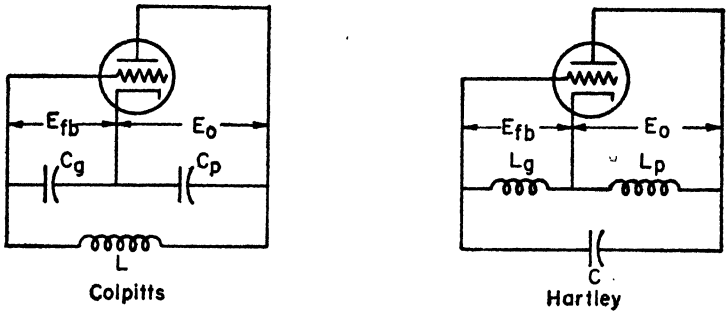


Fig. 8.9. Prototype oscillator circuits.

The conditions for oscillation may be established rather quickly for the ideal cases by application of the Barkhausen criterion. The equivalent plate circuits are shown in Fig. 8.10. Designating the tank circuit impedance across the terminals indicated as Z_L, for either circuit, then it is evident from elementary amplifier theory that the amplification in either case is

$$K_T = -\mu \frac{Z_L}{Z_L + r_p} = -\mu \frac{1}{1 + \left(\frac{r_p}{Z_L}\right)}$$

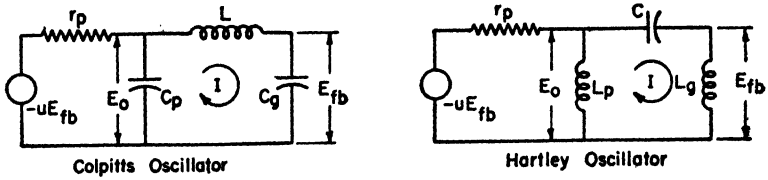


Fig. 8.10. Equivalent plate circuits of basic oscillator circuits.

Neglecting all losses, the input impedance of the tapped resonant circuits will be infinite and the voltage amplification of the stage approaches the amplification factor of the tube. That is,

$$K_T = -\mu$$

The feedback factors are obviously

$$\beta = -\frac{C_p}{C_g} \text{ for the Colpitts} \quad \beta = -\frac{L_g}{L_p} \text{ for the Hartley}$$

The Barkhausen criterion for each circuit is then

$$\mu = \frac{C_g}{C_p} \text{ for the Colpitts} \qquad \mu = \frac{L_p}{L_g} \text{ for the Hartley}$$

In practical cases the deviations from these relationships may be appreciable because of tank circuit losses and oscillator loading.

8.13 Single-tube UHF Oscillators

The necessity of using very small values of L and C in the tuned circuits of ultrahigh frequency oscillators yields two by-products. Due to the inevitable presence of the interelectrode capacitances from grid and plate to cathode, which are of about the size required for the tuned circuit, Colpitts oscillators are more common than the Hartley type. Secondly, the difficulty of obtaining lumped inductances (and capacitances) with values as small as required, and with the power-handling capacity desired, necessitates the use of resonant transmission lines as discussed in Chap. 5. Because of these facts, the physical appearance of a UHF oscillator may not seem to resemble either the Hartley or the Colpitts, but when reduced to its functional form, its true nature is revealed.

For example, consider the lighthouse tube oscillator of Fig. 8.5 which has been used as a local oscillator in UHF superheterodyne receivers. Only essential *RF* components are shown. The *RF* equivalent circuit may be developed in the following manner. The short-circuited coaxial transmission lines between cathode and grid, and

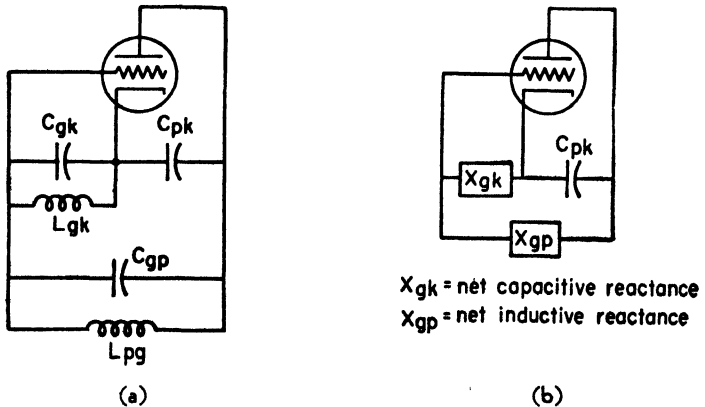


Fig. 8.11. (a) Direct equivalent RF circuit of lighthouse tube oscillator and tuner of Fig. 8.5. (b) Reduced RF equivalent circuit of lighthouse tube oscillator.

grid and plate, may be represented by inductances L_{gk} and L_{gp}, since they are assumed to be slightly less than three-fourths of a wavelength long. The quarter-wave choke (or *RF* choke, whichever is used) is an *RF* open circuit so that all elements connected to it may be neglected. Replacing the lines by inductances, as mentioned, and showing the tube interelectrode capacitances, the equivalent circuit of the lighthouse tube appears as shown in Fig. 8.11(a). The circuit may be rearranged somewhat by replacing the cathode circuit elements (L_{gk}, C_{gk}) and grid-to-plate circuit elements (L_{gp}, C_{gp}) by unknown reactances X_{gk} and X_{gp}. From the article on tapped resonant circuits it is known that the reactance on either side of the tap must be of the same type if feedback in the proper phase is to occur. Consequently, since the element between plate and cathode is a capacitance, it follows that the grid-to-cathode reactance must likewise be capacitive. Moreover, if this is true, and if a resonant circuit is to be formed, then the grid-to-plate reactance must be inductive. The final *RF* equivalent circuit appears as shown in Fig. 8.11(b) and is obviously a Colpitts oscillator. This is the general technique of analyzing UHF oscillators.

Tuning the grid-to-plate line determines the magnitude of the inductance L_{gp} which is the principal factor controlling the tuning of the oscillator to the desired frequency. The grid-cathode line determines the amount of feedback by establishing the value of X_{gk}. It has only a negligible effect upon frequency because the frequency of oscillation is

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{1}{L_{gp}C_0}}$$

where $$C_0 = C_{gp} + \frac{C_{gk}C_{pk}}{C_{gk} + C_{pk}} \approx C_{gp} + C_{pk}$$

since $$C_{gk} \gg C_{pk}$$

Another typical single-tube UHF oscillator is the ultra-audion shown in Fig. 8.12. Its *RF* equivalent circuit may be developed in the same general way. The condenser C_g is a d-c open circuit, but an *RF* short. It provides the proper bias by grid leak action and can be omitted from the *RF* equivalent circuit. The chokes in the cathode leads are *RF* open circuits and may be neglected in the analysis. R_p is a parasitic suppressor (see Art. 8.15) and R_g is the grid leak. Both may be omitted from the *RF* equivalent circuit.

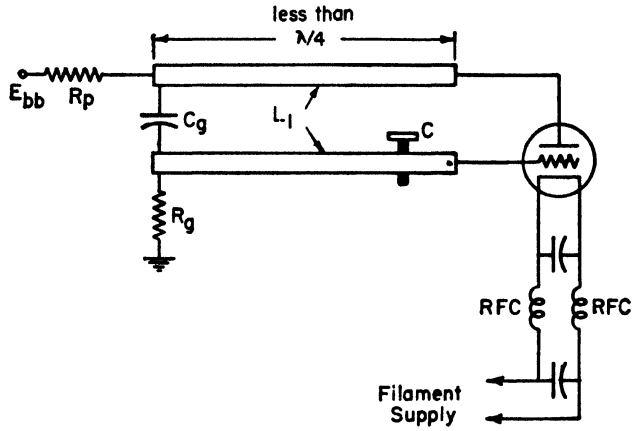
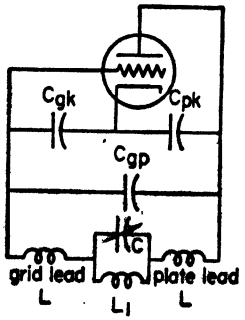


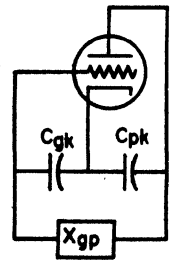
Fig. 8.12. Ultra audion UHF oscillator.

The transmission line, which is shorted at *RF* by C_g, is less than a quarter wavelength long and, accordingly, can be replaced by an inductance L. The screw, C, is simply a variable capacitance used for tuning purposes. Lumping the lead inductances and including the interelectrode capacitances yields the *RF* equivalent circuit of Fig. 8.13(a). Collecting the circuit elements in the grid-to-plate circuit into some unknown reactance X_{gp} yields the circuit shown in Fig. 8.13(b). This again is a Colpitts oscillator and it is apparent that X_{gp} must be inductive if feedback in proper phase is to be obtained.



RF Equivalent circuit of ultra audion oscillator

(a)



Reduced RF Equivalent circuit of ultra audion oscillator

(b)

Fig. 8.13.

8.14 Push-pull Oscillators

Any of the conventional UHF push-pull oscillator circuits such as the tuned-plate tuned-grid (TPTG), tuned-grid tuned-cathode (TGTK), or tuned-plate tuned-grid tuned-cathode (TPTGTK), may be simplified into prototype Hartley or Colpitts oscillators. They are similar in every respect to low-frequency oscillators except that resonant transmission lines are used in place of ordinary tuned circuits. Furthermore, lead inductances and interelectrode capacitances must be included in the analysis.

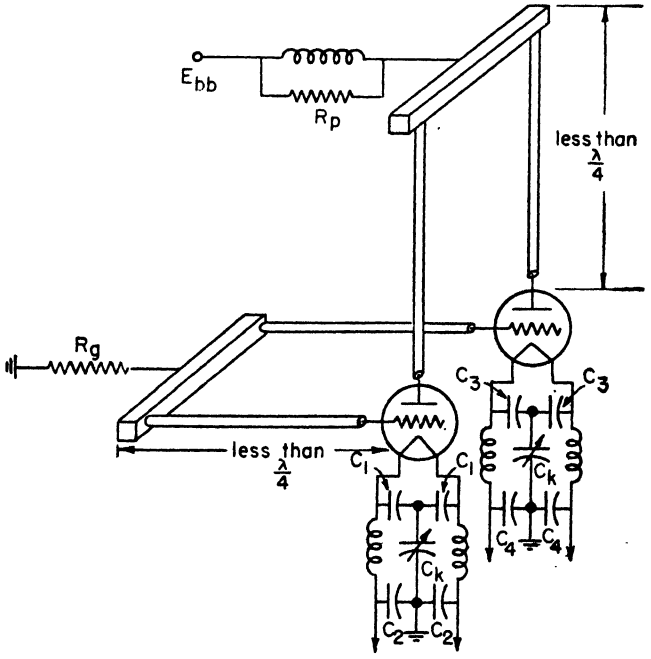


Fig. 8.14. Tuned-plate tuned-grid oscillator.

For example, consider the tuned-plate tuned-grid push-pull oscillator of Fig. 8.14. The by-pass condensers C_1, C_2, C_3, and C_4 appear as short circuits in the *RF* equivalent circuit, whereas the chokes appear as open circuits. The grid leak R_g and parasitic suppressor may both be considered as open circuits. The plate and grid lines are both less than a quarter wavelength and short-circuited so they may be represented by effective inductances L_p and L_g. The

plate and grid lead inductances are lumped in with L_p and L_g. The center points of these inductances are at *RF* ground potential because of the push-pull, balanced-to-ground connection. The variable capacitors C_k tune the cathode lead inductances to resonance, effectively grounding the cathodes. The *RF* equivalent circuit then appears as shown in Fig. 8.15(a).

Since the circuits associated with each tube are identical, analysis of a single tube will suffice for the whole circuit. Considering a single tube, and including the interelectrode capacitances, the *RF* equivalent circuit is shown in Fig. 8.15(b). Replacing the tuned circuits ($L_{gk}C_{gk}$ and $L_{pk}C_{pk}$) by unknown reactances yields the circuit of Fig. 8.15(c). For resonant transformer action to occur, it is obvious that both of these reactances must be inductive, thus producing a Hartley oscillator. This is the method of analysis best used for UHF push-pull oscillators.

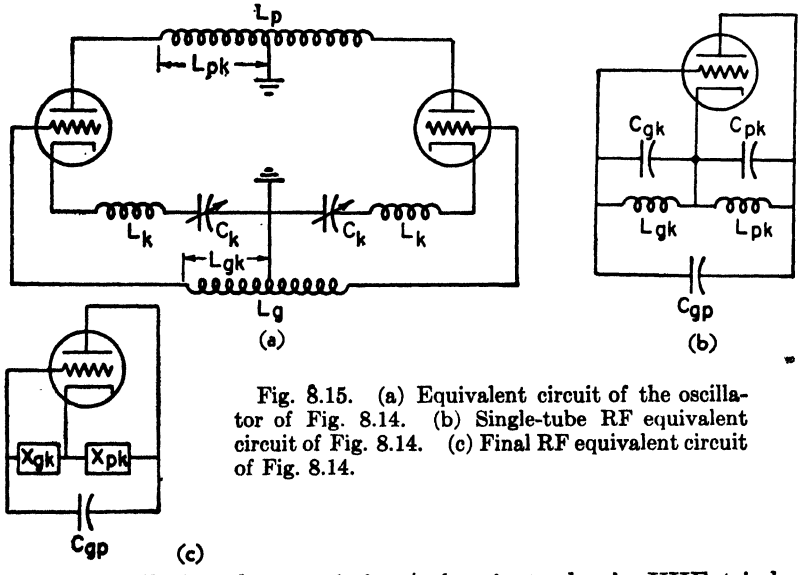


Fig. 8.15. (a) Equivalent circuit of the oscillator of Fig. 8.14. (b) Single-tube RF equivalent circuit of Fig. 8.14. (c) Final RF equivalent circuit of Fig. 8.14.

The small size characteristic of the electrodes in UHF triodes makes it very difficult to obtain large amounts of power, even from push-pull circuits. Consequently, more tubes are frequently required. The resulting circuit is called a *ring oscillator* because several push-pull units are combined around a ring and made to operate as a single circuit. Such networks always contain an even number of

tubes because they are made up of sets of push-pull units. They are arranged in such a way that adjacent tubes around the ring are always in opposite phase. Therefore, any two adjacent tubes form a push-pull circuit. Further discussion of the circuit is not warranted because the ring oscillator has the same fundamental characteristics as the push-pull unit already discussed. The same method of analysis is used, regardless of the number of tubes employed in the ring.

8.15 Parasitics and Their Suppression

The essential components of an ultrahigh frequency, tuned-plate tuned-grid push-pull oscillator are shown in Fig. 8.16. Since the tubes operate in push-pull, at some instant of time the plates and grids must have the relative polarities indicated. There are standing waves on the grid and plate lines due to their resonant

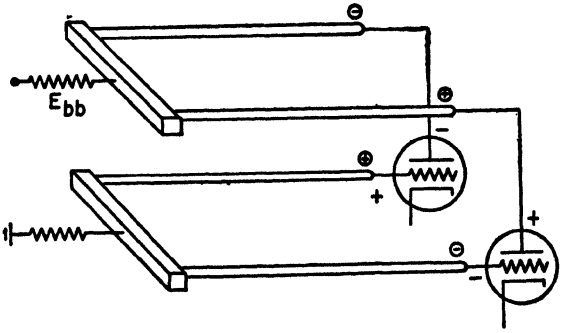


Fig. 8.16. Oscillator circuit susceptible to generation of parasitics. Resistors are parasitic-suppressors.

characteristics. A resonant line *may* be formed between the grid and plate line of each tube due to the open-circuited termination, or, if the lines are by-passed to ground, due to the resulting short-circuited termination. This would tend to cause oscillation at another frequency in a wrong mode. It could set up in-phase standing waves and the tubes might no longer operate push-pull. Unwanted oscillations in improper modes are called *parasitics*.

Parasitics can occur in a great many ways. The method given above is both typical and common, that is, by the formation of resonant lines between two conductors that are individually parts of other circuits. Metallic shields between conductors are seldom of assistance because of the *images* formed, which may cause parasitics to occur between the conductor and its image.

Parasitics, which are generally at frequencies other than that desired in the main mode, should be prevented because they lower the efficiency of the system and are often illegal if they have sufficient strength to radiate. Moreover, they do not necessarily respond to the tuning adjustments of the desired mode.

Wrong modes may be suppressed by keeping the parasitic lines from being resonant. This can be accomplished by terminating these lines in their characteristic impedance in such a way that operation in the main mode is unaffected. One method of doing this for the circuit of Fig. 8.16 is shown by the location of the parasitic-suppressing resistors. Unfortunately, it is seldom possible to specify exact values for these resistances, so that they are usually experimentally determined. Their values ordinarily range from 10 to 100 ohms in transmitters, and perhaps several thousand ohms in receiver local oscillators.

Careful physical arrangement of the oscillator components will aid materially in the suppression of parasitics and should be used in preference to resistors whenever possible.

8.16 Power Supply Connections

Many ultrahigh frequency oscillators require large *RF* voltages at the filaments, measured with respect to ground. However, *RF* voltages are extremely undesirable at the source of filament power. A particularly good reason for not wanting *RF* across the filament transformer arises from dielectric considerations. At the frequencies normally encountered, the dielectric material used for insulating the transformer windings is not very good. The presence of appreciable *RF* voltages would cause large currents to flow, heating the dielectric, and possibly breaking it down. In any case, the *RF* power loss in the transformer is undesirable, making a real dent in the efficiency of the system.

At ultrahigh frequencies, the lead inductance of the filament circuit has a relatively high reactance so that the *RF* currents passing through the leads produce large *RF* voltage drops. Unless special measures are taken, this high *RF* voltage to ground appears across the filament transformer producing the effect noted in the preceding paragraph.

At comparatively low frequencies, the *RF* voltage at the filament may be shorted to ground with by-pass condensers without disturbing the d-c filament current path. This is not too satisfactory, however,

because the condensers are required to pass large *RF* currents which would tend to introduce the always undesirable dielectric losses. A somewhat better arrangement would be to place *RF* chokes in series with the filament leads. They have practically no effect upon the d-c path, but they have a very high reactance at the oscillator frequency so that most of the drop in *RF* voltage occurs across them, leaving little to appear across the filament transformer. This could be improved even further by inserting by-pass condensers to ground near the filament transformer, effectively removing whatever remains of the *RF* voltage.

A further refinement in the system would be to add a by-pass condenser *between* filament leads at the filament end of the chokes. This tends to equalize the *RF* voltages in the two leads, removing the possibility of an *RF* field between them.

At higher frequencies, lumped-circuit constants are generally impractical and are replaced by resonant transmission lines. A

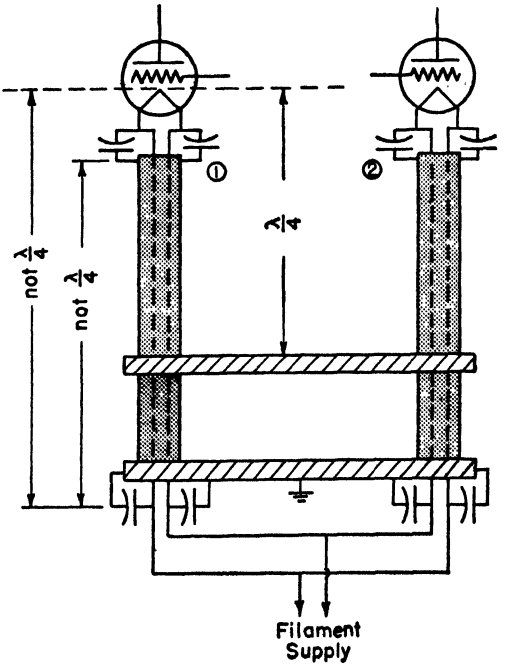


Fig. 8.17. Typical filament supply connections for UHF oscillators.

typical arrangement is shown in Fig. 8.17. The *RF* voltage to ground, developed in the filament leads, is by-passed by the condensers at points (1) and (2) over to the outside of the hollow conductors. The *RF* field to ground may be considered as existing between the outer conductor and its image in the ground. The end of this line is shorted to ground. Since it is short-circuited and a quarter of a wavelength long, it appears as an *RF* open circuit. Similarly, any field between the two conductors encounters an open circuit. Consequently, negligible *RF* voltages appear at the filament transformer.

Similar techniques are employed in isolating the plate power supply from the *RF* energy. *RF* chokes in series with the power supply leads are common, as well as the use of by-pass condensers to ground. At ultrahigh frequencies, quarter-wave sections of transmission line are used which ideally have infinite input impedance to *RF*. In such cases they simply replace the *RF* chokes and are quite appropriately called quarter-wave chokes.

8.17 Output Coupling

After the oscillator has generated the required signal, some means of removing the energy and putting it on a transmission line must be provided so that it can fulfill its intended purpose. Coupling may be obtained by any one of the three fundamental methods:

(1) Conductive coupling — direct coupling.
(2) Inductive coupling — through the magnetic field.
(3) Capacitive coupling — through the electric field.

Besides merely removing energy, the coupling device must match the transmission line and oscillator in such a way that the proper resistance is introduced into the oscillating circuit and the line is terminated in its characteristic impedance.

Direct coupling can be used only when the transmission line leading to the antenna is a parallel wire line, balanced to ground, like the tuned lines in the oscillator. Furthermore, there cannot be any d-c voltages on the tuned lines to which the transmission line is connected. Therefore, conductive coupling is limited in its application to circuits like the tuned-plate tuned-cathode push-pull oscillator.

When the tuned lines carry high d-c voltages, such as in the tuned-plate tuned-grid oscillator discussed in Art. 8.14, conductive

coupling is not feasible. In such circuits, coupling is best accomplished through the medium of the magnetic field. In technique, it takes the form of a coupling loop located at a point where the magnetic field is a maximum. In many cases the loop is tuned by a variable capacitance, usually in the form of a screw. The transmission line is connected to the loop by means of taps that are generally adjustable in position on the loop. The circuit behaves in very nearly the same manner as an ordinary air core transformer and hardly warrants additional discussion.

Capacitive coupling is frequently used in oscillators employing coaxial lines. The coupling device usually takes the form of a probe inserted into the line at a point of maximum electric field intensity. This technique was mentioned in Chap. 7 in connection with waveguide excitation and is used in the same fashion with oscillators. Although commonly used with lighthouse tube oscillators, it is not as widely employed as inductive coupling.

8.18 Positive Grid Oscillators

The UHF oscillators discussed in this chapter up to the present point differ from ordinary low-frequency oscillators only in their general physical appearance. The internal characteristics and principles of operation are the same. The positive grid oscillator, very briefly discussed in this article, is an entirely different proposition and analysis of its operation proceeds upon radically different lines.

The phenomenon which permits oscillation in positive grid tubes was correctly identified as early as 1919 by Barkhausen and Kurtz who encountered it in the course of their investigations concerning the effect of residual gas in triodes. The circuit was used for a while in early experimental microwave systems, but because of the very low efficiencies obtainable, it is currently of little more than academic interest.

The positive grid oscillator is included in this book principally to indicate a method of analysis, which will be used in the next two chapters, that is based upon a consideration of the electronic trajectories rather than the associated external circuit. The approach is so fundamentally different from that ordinarily used in analyzing circuit behavior that it is best introduced by stages.

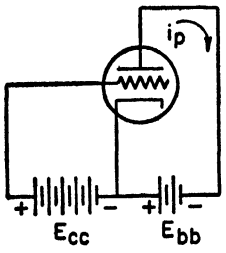


Fig. 8.18. Circuit in which current flow exists in the plate circuit when the plate is negative.

Consider the circuit diagram of Fig. 8.18. Observe that the grid is operated at a relatively high positive potential and the plate is held negative. Theoretically, it would appear that no current would flow in the plate circuit, neglecting any positive ion current due to residual gas. Yet, the experiment proves this false, because some current does exist in the plate circuit.

The answer to this apparent anomaly is obtainable from a consideration of the electrons, rather than an overhasty examination of the circuit connections. An electron at the cathode is accelerated by the positive voltage E_{cc} between grid and cathode. This electron may do either of two things: it may

(1) Strike a grid wire and be removed from the tube.
(2) Pass through an opening in the grid and enter the grid-plate space.

In the latter instance, the electron enters a retarding field, due to the voltage $(E_{cc} + E_{bb})$ between plate and cathode, which tends to return the electron toward the grid. Inasmuch as the kinetic energy acquired by the electron during acceleration is insufficient to overcome the effects of the decelerating field, the electron does not reach the plate, but is repelled and falls back toward the grid. As before, it either strikes a grid wire and is removed, or it passes through an opening in the grid and returns to the space charge cloud near the cathode. It is evident that some few electrons may make several oscillations back and forth through the grid. Because of the repelling action of the plate, the plate power supply is required to deliver power and develops a current flow as a consequence.

For effective operation as an oscillator, an alternating field must be introduced in such a way that the electrons deliver more energy to this field by being decelerated than they receive by being accelerated. There are an indefinite number of ways of introducing tuned circuits to accomplish this, but discussion of one should suffice for the purpose of illustration.

Consider the circuit of Fig. 8.19. Assume that by some means, the

tuned grid circuit is excited into oscillation so that the voltage e_g is sinusoidal. The total voltage on the grid accelerating the electrons is then $(E_{cc} + e_g)$. If an electron leaves the cathode at a time when the alternating component of grid voltage is going through zero in the negative direction, and if the frequency of the alternation equals the electron transit time from cathode to point of rest near the plate before returning through the grid again, then the electron does work against the a-c field in going from cathode to grid. Ideally, this electron would pass through the grid when the voltage is going through zero in the positive direction, and, on the outward trip toward the plate, it would again be working against the a-c field, delivering energy to it.

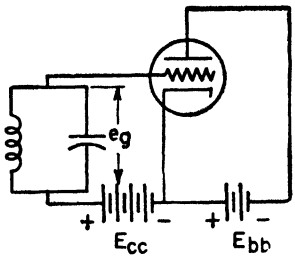


Fig. 8.19. Typical positive-grid oscillator circuit.

On the return trip, it is evident that this "favorable" electron will continue to deliver energy to the a-c field and it is seen that the individual electron functions as the converter of d-c accelerating energy to a-c in the grid circuit. The electron may make several such excursions, but because of gradually diminishing kinetic energy, it will dance with less and less amplitude. Since the period of the electronic oscillation depends upon the amplitude of its motion, the electron will eventually get out of step with the a-c grid voltage and start absorbing energy from the alternating field. This is a contributing factor to the low efficiency of the circuit.

On the other hand, an electron leaving the cathode, when the grid voltage is going through zero in the positive direction, will receive energy from the a-c field and be further accelerated. By proper adjustment of the plate voltage, relative to the amplitude of the alternating grid voltage, this increase in electronic kinetic energy can be made sufficiently large to cause the electron to strike the plate, thus removing these "unfavorable" electrons after a single excursion through the grid.

Inasmuch as the favorable, or energy-delivering, electrons make several excursions through the grid, while the unfavorable, or energy-removing, electrons generally make only a single traverse, there is a net transfer of energy to the grid circuit. This energy can be made

sufficiently large to regenerate the original disturbance postulated and make the system oscillate.

REFERENCES

Alfven, H., "The Barkhausen-Kurtz Generator," *Phil. Mag.*, **19**, 419 (1935).

Anderson, J. E., "Theory of Electron Oscillators," *Electronics*, **9**, 9 (1936).

Englund, C. P., "The Short Wave Limit of Oscillators," *Proc. IRE*, **15**, 914 (1927).

Ferris, W. R., "The Input Resistance of Vacuum Tubes as UHF Amplifiers," *Proc. IRE*, **24**, 82 (1936).

Gavin, M. R., "Triode Oscillators for Ultra Short Wave Lengths," *Wireless Eng.*, **16**, 287 (1939).

King, R., "Beam Tubes as UHF Generators," *J. Applied Phys.*, **10**, 638 (1939).

Lindenblad, N. E., "Development of Transmitters for Frequencies Above 300 mcps," *Proc. IRE*, **23**, 1013 (1935).

Llewellyn, F. B., "Equivalent Networks of Negative Grid Vacuum Tubes," *Bell System Tech. J.*, **15**, 4 (1936).

Electron Inertia Effects, Cambridge Univ. Press, London, 1941.

Samuel, A. L., "Extending the Frequency Range of the Negative Grid Tube," *J. Applied Phys.*, **8**, 677 (1937).

"A Negative Grid Triode Oscillator and Amplifier for UHF," *Proc. IRE*, **25**, 1243 (1937).

Samuel, A. L., and Sowers, N. E., "A Power Pentode For UHF," *Proc. IRE*, **24**, 1464 (1936).

Seeley, S. W., and Anderson, E. I., "UHF Oscillator Frequency Stability," *RCA Rev.*, **5**, 77 (1940).

CHAPTER 9

KLYSTRONS

THE preceding chapter has dealt with the adaptation of conventional vacuum tubes for use in the ultrahigh frequencies. The transit time of the electron from cathode to anode was shown to be one of the chief obstacles to effective use of these tubes for UHF applications. It was evident that an upper frequency limit exists beyond which further refinement of conventional tubes appears to be impractical. Rather than making a direct assault on the problem by trying to overcome the disadvantages arising from transit time, new tubes have been developed which put this apparent drawback to work by making operation depend upon the time of flight of the electron from cathode to anode. At the present time, these "transit time tubes" fall into two general classes:

(1) Velocity-modulated tubes (covered in this chapter).
(2) Magnetrons (covered in the next chapter).

The most predominantly successful tube in current use which employs the principle of velocity modulation of an electron beam is the Klystron*. It was developed in 1939 by the Varian brothers and found an extensive application very rapidly during the Second World War in military radar. Due to certain operating characteristics, it is currently finding a definite use in microwave relay stations in conjunction with television network development. Due to its comparative ease of tuning, relative to other microwave oscillators, it is widely used as a variable-frequency UHF signal generator and as a local oscillator in microwave superheterodyne radio receivers.

The rather detailed discussion of the Klystron presented in this chapter is justified by the fact that it is representative of velocity-modulated tubes in general. Other tubes of the same character are

*Sperry Gyroscope Company trade-mark

listed for the purpose of general information, but further discussion of them does not appear to be warranted because of the basic similarities of operation.

9.1 Velocity Modulation

In many vacuum tube applications, the primary function of the vacuum tube as a circuit element is to convert d-c energy into a-c. Thus, the fundamental requirement which must be met by transit time tubes is that the moving electrons must transfer some of their kinetic energy, acquired from a d-c supply, to a source supplying an alternating field. This energy transfer is accomplished by passing electrons through a retarding electric field. More specifically, these electrons must deliver more energy to this source by being decelerated than they remove by being accelerated. In other words, for effective operation, more electrons must pass through the alternating field when it is exerting a retarding action than when it is producing an accelerating force. Statistically, such an unequal division does not occur, so that as long as the electron stream is undisturbed, no net transfer of energy will occur. In effect then, the problem of using transit time to advantage revolves around the necessity of increasing the *number* of electrons doing work against the alternating field, or increasing the *amount* of work that each of the favorable electrons does against this field, or both.

A commonly used method of accomplishing the requirement listed above is called *velocity modulation*, a technique whereby the velocity of the electron in the beam is modulated by some specified electrode voltage. In other words, the velocity variations of the electrons in the beam are controlled by proportional variations in the modulating-signal voltage. The purpose of velocity modulation is to make most of the electrons pass through the alternating field at a time when they can deliver the most energy.

The velocity of an electron beam can be modulated by the simple electrode configuration shown in Fig. 9.1. In essence, the structure consists of two adjacent grids through which the electrons must pass after being accelerated by the steady electric field between the cathode and hollow cylindrical anode. Assuming that the steady accelerating field produces a large electron velocity, then if the grid spacing is small, the electrons pass between the grids in a very short period of time. As a matter of fact, the grid spacing and d-c field are generally

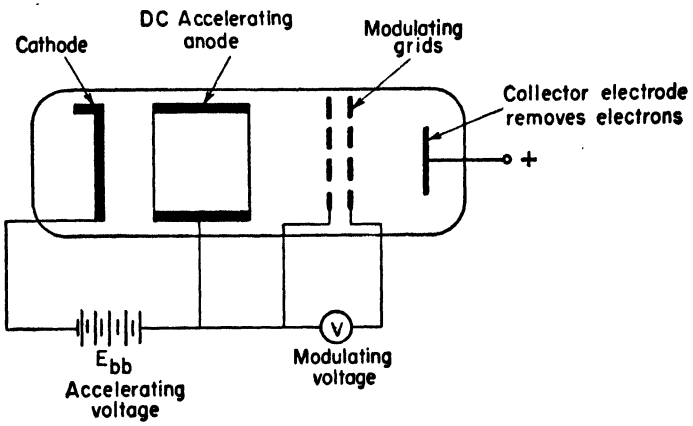


Fig. 9.1. Basic velocity-modulated tube.

adjusted to such values that the transit time of the electrons between grids is a very small fraction of the period of the alternating grid voltage. This grid voltage is the velocity-modulating field, *not* the field used for energy removal. Electrons emerging from the grid structure have velocities that depend upon the instantaneous sum of the d-c field and the field between the velocity-modulating grids.

There are three distinguishable cases depending upon the character of the alternating field while the electron is between grids. The first, and most obvious, case is when the voltage is zero. The electron traverses the grid space and emerges with its velocity unchanged. Electrons entering when the modulating field is decreasing are decelerated, whereas they are accelerated if the field between grids is increasing. As a result, the velocity of the electron is a function of the grid voltage and the electron beam has been velocity-modulated.

If it is assumed that the emission of electrons from the cathode occurs at equal intervals and with equal velocities, then—since the field in the space between the cathode and first grid is uniform—at this grid the number of electrons, their velocities, and the spacing between them are all unvarying with respect to time. Furthermore, since the grid spacing is very small, and because the d-c field is ordinarily much larger than the modulating field between the grids, the number and spacing of electrons emerging from the second grid are uniform. However, their velocities are different, having been modulated by the grid signal.

Ideally then, the modulating field receives as much energy by decelerating the electrons as it delivers by accelerating them. This would appear to indicate that zero power is required for velocity modulation. This condition is never quite met in practical use, but the power is generally smaller than the losses in the tuned circuit associated with the modulating grids and can be neglected without substantial loss of accuracy.

For greater effectiveness, the simple cathode structure assumed in Fig. 9.1 is quite generally replaced by an electron gun of the type customarily found in cathode ray tubes.

9.2 Conversion to Intensity Modulation

After modulating the velocity of the electron beam as outlined in the preceding article, it becomes necessary to establish some method of deriving a useful output from the tube. This cannot be done by the simple combination of electrodes shown in Fig. 9.1 because velocity modulation by itself does not increase the number of electrons working against the retarding field, nor does it increase the net amount of energy delivered. The necessary prerequisite for operation is to convert the velocity-modulated electrons into an intensity-modulated beam. An electron beam constitutes a convection current and when that current is made to vary as a function of time, the beam is said to be intensity-modulated.

There are several methods of converting from velocity to intensity modulation, but the most widely used is the *drift* tube. The *drift space* is merely an equipotential region in which the velocity-modulated electrons, leaving the modulating grids, may coast freely with the velocities imparted to them by the modulating voltage. Due to the variations in electron velocities produced by the modulating voltage, the electrons tend to cluster together after drifting for some distance in this equipotential space. Thus, the beam becomes intensity-modulated as the electrons begin to form into clusters or bunches. In other words, the electron density, or the charge intensity, has been made to vary in a manner roughly proportional to the modulating signal.

9.3 The Two-cavity Klystron

Even after conversion to intensity modulation, the elementary tube discussed in the preceding articles is still impractical because

there is no provision for energy extraction from the intensity-modulated electron beam. The required energy removal is most readily accomplished by inserting a second pair of grids in the drift tube at a point where the electrons are well bunched together. Then, by applying an electric field between these grids that exerts a maximum retarding force when the electron bunches pass through, the clusters of electrons are decelerated and their loss in kinetic energy is transferred to the source producing the retarding field.

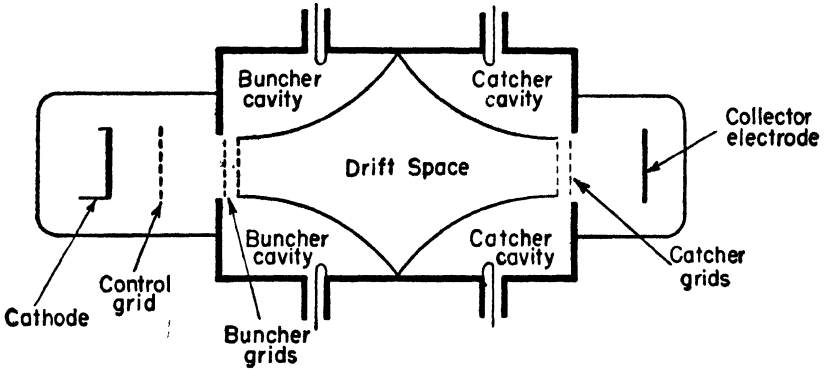


Fig. 9.2. Two-cavity system.

A common tube working on this principle is the *two-cavity Klystron*. The cavities are the tuned circuits associated with each pair of grids. A schematic sketch of this tube is shown in Fig. 9.2. The Klystron operates in the following manner. The electrons are formed into a uniform beam by the electron gun. The electrons are shot through the *buncher* grids, which velocity-modulate the beam. The electrons emerge from the buncher grids with varying velocities, which, as they traverse the drift space, cause the electrons to cluster into bunches forming an intensity-modulated beam. These bunches pass through the *catcher* grids at a time when the catcher voltage exerts maximum opposition, thus removing energy from the electrons. After passing through the catcher the electrons are picked up by the *collector* electrode and removed from the tube.

9.4 Applegate Diagram

A graphical representation that shows the electron position inside the Klystron as a function of time is called an *Applegate diagram*. It is a definite aid to an understanding of the operation of transit time

tubes and assists in deducing certain characteristics that are not immediately obvious from a discussion. A complete Applegate diagram for a two-cavity Klystron appears in Fig. 9.3.

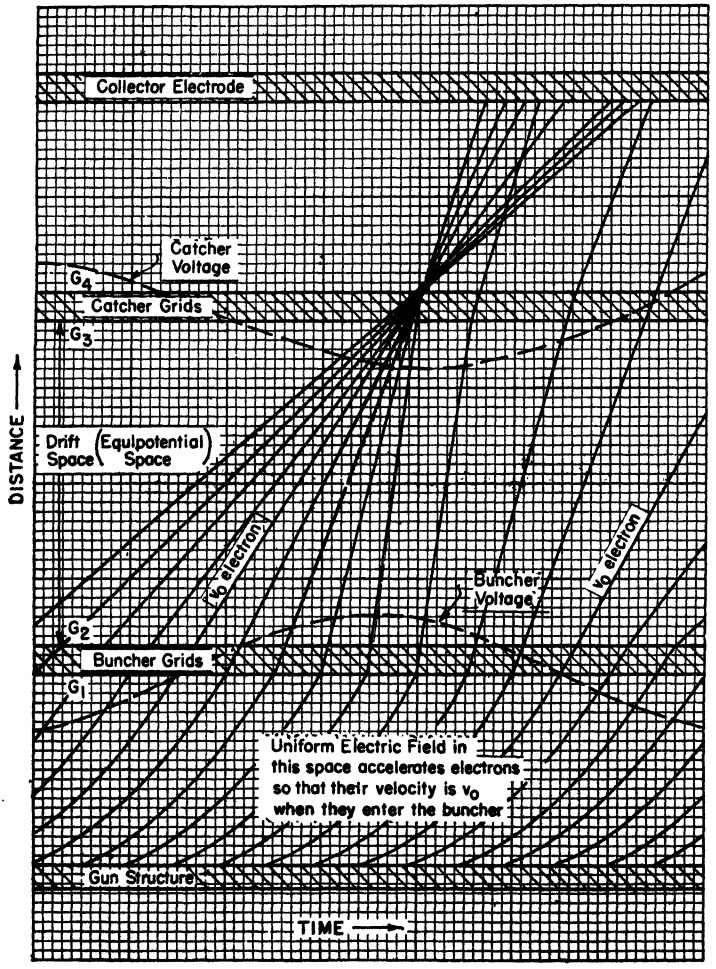


Fig. 9.3. Applegate diagram.

The solid lines in Fig. 9.3 represent the positions of the individual electrons at any point inside the Klystron as a function of time. Scrutiny of the electron starts as it emerges from the electron gun.

It will be noted that it has been assumed that the electrons leave the electron gun at equal intervals of time. Furthermore, since the slopes of these lines are proportional to the electronic velocity, it is also apparent that it has been assumed that the electrons have equal velocities when they leave the gun structure.

In the region between the gun and the first of the buncher grids, the electrons are continuously accelerated, so that the velocity increases uniformly as shown by the curvature of the lines. Since all electrons are acted upon by the same force, they all have the same velocity, v_0, when they enter the buncher grids. The voltage on the buncher grids, $E \sin \omega t$, is indicated by the dotted-line sine wave.

If the alternating voltage on the buncher is zero when the electrons pass through, their velocity remains v_0 and they fall through the drift space with this velocity. They are called v_0 *electrons*. If the buncher voltage is positive, the electrons are speeded up (steeper slope) and tend to overtake the v_0 electrons in the drift space. If the buncher voltage is negative, the electrons are slowed down and the v_0 electrons tend to catch up. Consequently, the buncher voltage causes the electrons to cluster about the v_0 electrons. The place in the drift space where the electrons are most closely clumped together is called the *point of maximum bunch* (PMB).

The catcher grids are generally located close to the PMB as shown on the Applegate diagram. The catcher grid voltage is shown by the dotted sinusoidal wave in the figure. It is adjusted, as shown, so that the bunch passes through when the voltage is at a negative maximum so that it is exerting maximum retarding force.

9.5 Bunching

The space charge represented by the concentration of electrons into bunches tends to destroy the bunch formation because of the mutual forces of repulsion between electrons. The effect operates to make the bunches less sharply defined and it is called *de-bunching*. Both longitudinal and transverse de-bunching may occur and both tend to decrease the tube efficiency.

When the bunch is formed after the electrons pass through the catcher, the beam is said to be *under-bunched*. The magnitude of the buncher grid voltage should be increased to improve the bunching process. Conversely, if the point of maximum bunch occurs before the electrons reach the catcher, the beam is *over-bunched* and the

velocity-modulating voltage should be decreased. *Optimum* bunching occurs when the bunch forms at the center of the catcher grids.

9.6 Catcher Current

The current in the catcher may be written directly as follows:

$$i_c = nev = \text{catcher current}$$

where n is the number of electrons in the catcher as a function of time, e is the charge on the electron, and v is the electron velocity, which is also a function of time. The electron velocity varies in direct proportion to the buncher voltage, so the instantaneous electron velocity is composed of a constant term (v_0) and a sinusoidal term. Consequently, the catcher current is

$$i_c = ne(v_0 + v \sin \omega t)$$

Generally speaking, the velocity imparted to the electrons by the d-c accelerating field is very much larger than the velocity changes produced by the buncher grid voltage. Hence,

$$v_0 \gg v$$

so that

$$i_c = nev_0$$

Moreover, since v_0 and the electronic charge are constants, the catcher current is practically proportional to the number of electrons in the catcher at any given instant. Therefore, the current at any point in the Klystron as a function of time could be obtained from the Applegate diagram by plotting the quantity of electrons crossing any horizontal line drawn on the chart.

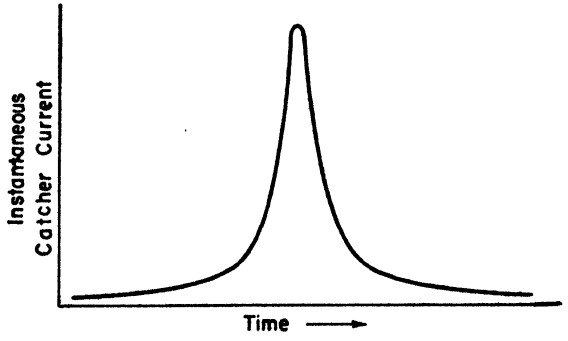


Fig. 9.4. Catcher current at the point of maximum bunch.

Following this procedure across a line which passes through the point of maximum bunch yields the rather idealized catcher current waveform, shown in Fig. 9.4, where the catcher was assumed to be located at the PMB. It will be noted from a careful consideration of the Applegate diagram that, slightly beyond the PMB, the electron beam splits into two smaller bunches. Consequently, if the catcher were located at such a point, the catcher current would be as in Fig. 9.5. This double-peaked response is characteristic of the catcher current in actual operation of a practical Klystron. The high harmonic content of such a waveform suggests using the Klystron as a frequency multiplier.

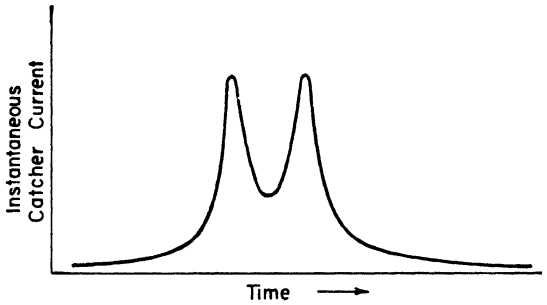


Fig. 9.5. Catcher current at a point beyond the point of maximum bunch.

9.7 Phase Shift

Phase shift, or time delay, is essential to the operation of a Klystron since its use depends upon the transit time in the drift space. This is certainly a different viewpoint from that associated with ordinary tubes.

There are three separate sources of time delay in a Klystron that may be discussed separately:

(1) Transit time.
(2) The mechanism of bunching.
(3) The feedback circuit.

The last named source is not present when the Klystron is used as an amplifier, but it is very important when operated as an oscillator. Let

θ_1 = phase shift due to transit time
θ_2 = phase shift due to bunching
θ_3 = phase shift due to the feedback circuit

Phase shift due to transit time. If the frequency of the buncher voltage is f, and if the time of flight of the electron between buncher and catcher is t, then the phase shift due to transit time is

$$\Theta_1 = 2\pi ft \tag{9.1}$$

The transit time t is the drift distance, s, divided by the electron velocity, or

$$t = \frac{s}{v_0 + v \sin \omega t} \tag{9.2}$$

However, as was pointed out in the preceding article,

$$v_0 \gg v$$

and consequently, the transit time is approximately

$$t = \frac{s}{v_0} \tag{9.3}$$

Therefore, the phase shift due to this transit time is

$$\Theta_1 = 2\pi f \frac{s}{v_0} \tag{9.4}$$

The velocity of the electron may be evaluated by equating the electronic kinetic energy to the energy delivered by the electric field. That is,

$$\tfrac{1}{2}m(v_0 + v \sin \omega t)^2 = e(E_0 + E \sin \omega t) \tag{9.5}$$

However, as a general rule, as previously indicated,

$$v_0 \gg v \qquad \text{and} \qquad E_0 \gg E$$

Therefore,

$$\tfrac{1}{2}mv_0^2 = eE_0 \tag{9.6}$$

Then, solving for the electron velocity, v_0, yields

$$v_0 = \sqrt{2\left(\frac{e}{m}\right)E_0} \tag{9.7}$$

where e is the electron charge, m is the electron mass, and E_0 is the d-c accelerating field in the Klystron. Hence, the phase shift due to transit time is

$$\Theta_1 = 2\pi fs\sqrt{\frac{1}{2}\left(\frac{m}{e}\right)\frac{1}{E_0}} \tag{9.8}$$

Substituting the proper values for the electronic charge and mass yields

$$\Theta_1 = \left(\frac{\pi}{3}\right)\frac{fs}{\sqrt{E_0}}\,10^{-5}\text{ rad} \tag{9.9}$$

where s is the drift space in meters, f is the buncher voltage frequency in cycles per second, and E_0 is the d-c accelerating field in volts.

The relative magnitude of the phase shift due to transit time may be determined from the constants of a typical 10-cm Klystron, for which

$$s = 0.0254\text{ m}$$
$$f = 3 \times 10^9\text{ cps}$$
$$E_0 = 1000\text{ volts}$$

Substitution into Eq. (9.9) yields

$$\Theta_1 = 8\pi\text{ rad} = 4\text{ cycles}$$

Phase shift due to bunching. From an earlier discussion it will be recalled that effective operation of the Klystron depends upon the electron bunch passing through the catcher when the catcher field is exerting a maximum retarding force. The electrons tend to bunch about the v_0 electron that crosses the buncher when the buncher voltage is going through zero in the positive direction. Since, for proper operation, they arrive when the catcher voltage has its maximum negative value, there is an inherent 90° phase advancement associated with the mechanism of bunching. That is,

$$\Theta_2 = \frac{\pi}{2}\text{ rad} \tag{9.10}$$

This phase shift may be thought of as the difference that exists between the buncher and catcher voltages when the v_0 electron is used as the basis for comparison. Another way of looking at it is to observe that a current maximum in the catcher is associated with zero field in the buncher.

Phase shift in the feedback circuit. Feedback between catcher and buncher cavities is generally provided by means of a small coaxial cable. The phase shift in such a cable may be written as

$$\Theta_3 = 2\pi f t_3 \tag{9.11}$$

where t_3 is the time of transit in the cable. But,

$$t_3 = \frac{l}{v_c} \tag{9.12}$$

where l is the effective length of the cable and v_c is the velocity of transmission in the cable. Therefore, the phase shift due to feedback is

$$\Theta_3 = \frac{2\pi fl}{v_c} \tag{9.13}$$

Total phase shift. When the Klystron is operated as an amplifier, the total phase shift is

$$\Theta_T = \Theta_1 + \Theta_2 = \left(\frac{\pi}{3}\right)\left(\frac{fs}{\sqrt{E_0}}\right)(10^{-5}) + \frac{\pi}{2} \tag{9.14}$$

When operated as an oscillator, the phase shift is

$$\Theta_T = \Theta_1 + \Theta_2 + \Theta_3 = \left(\frac{\pi}{3}\right)\left(\frac{fs}{\sqrt{E_0}}\right)(10^{-5}) + \frac{\pi}{2} + \frac{2\pi fl}{v_c} \tag{9.15}$$

9.8 Klystron as an Oscillator

In the preceding article it was shown that the total phase shift in a Klystron connected as an oscillator is given by

$$\Theta_T = \Theta_1 + \Theta_2 + \Theta_3 \tag{9.16}$$

or, substituting the expressions derived in Art. 9.7,

$$\Theta_T = \frac{2\pi fs}{v_0} + \frac{\pi}{2} + \frac{2\pi fl}{v_c} \tag{9.17}$$

However, $$f = \text{frequency} = \frac{c}{\lambda}$$

where c = velocity of light and λ = wavelength. Hence, the expression for the total phase shift becomes

$$\Theta_T = 2\pi\left(\frac{sc}{v_0\lambda}\right) + \frac{\pi}{2} + 2\pi\left(\frac{lc}{v_c\lambda}\right) \tag{9.18}$$

But, for oscillations to occur, the total phase shift must be some multiple of 2π rad. Then, assuming this condition,

$$\Theta_T = 2N\pi \qquad \text{where } N = \text{any integer} \tag{9.19}$$

Equation (9.18) becomes

$$(2\pi)N = 2\pi\left(\frac{sc}{v_0\lambda} + \frac{1}{4} + \frac{lc}{v_c\lambda}\right) \tag{9.20}$$

Solve Eq. (9.20) for $(sc/v_0\lambda)$, obtaining

$$\frac{sc}{v_0\lambda} = N - \frac{lc}{v_c\lambda} - \frac{1}{4} \tag{9.21}$$

Using the equation for v_0 previously calculated yields

$$\frac{500s}{\lambda\sqrt{E_0}} = N - \frac{lc}{v_c\lambda} - \frac{1}{4} \tag{9.22}$$

Therefore, in order for oscillations to occur, the value of the d-c accelerating field, E_0, must be adjusted to a value that will satisfy this equation. The process of meeting the condition for oscillation is evidently one of adjusting the transit time to some particular critical value.

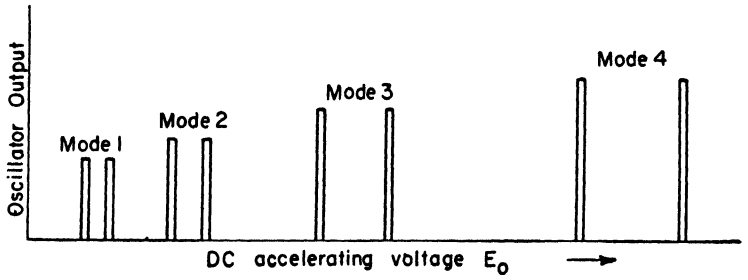


Fig. 9.6 Effect of d-c accelerating voltage E_0 on the oscillation characteristic of a two-cavity Klystron.

Consequently, it is apparent that only certain critical values of E_0 will reduce the total phase shift to some multiple of 2π, so the Klystron will oscillate only for particular values of this voltage. This is illustrated in Fig. 9.6. Generally, the higher the voltage, the stronger the oscillation.

The formula derived for $(sc/v_0\lambda)$ may be plotted as shown in Fig. 9.7. The slope is (λ/s). The intercept on the H axis is the term $(lc/v_c\lambda + \frac{1}{4})$. Thus, if this curve is obtained experimentally, a means is available for evaluating the phase shift in the feedback cable that couples the two cavities.

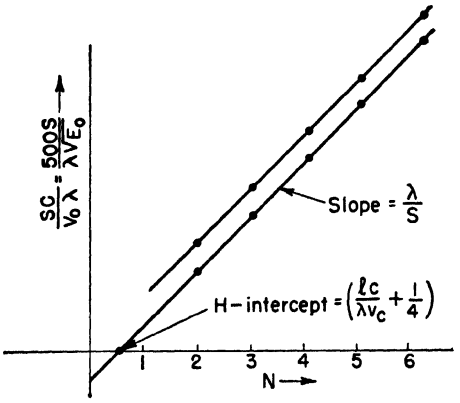


Fig. 9.7. Critical voltages E_0 for Klystron oscillation.

9.9 Modes of Oscillation

The two resonators associated with the buncher and catcher grids are very closely coupled together. It will be recalled that two identical tuned circuits that have more than the critical amount

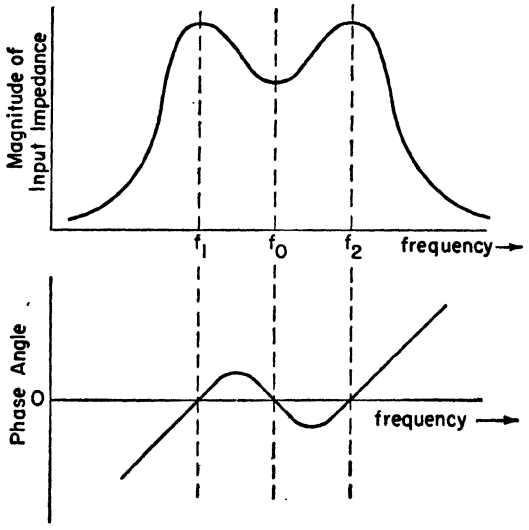


Fig. 9.8. Frequency characteristics of two overcoupled tuned circuits resonant at the same frequency.

of coupling are resonant at two distinct frequencies. The frequency characteristics for such a circuit appear as shown in Fig. 9.8. Observe that the phase shift is zero at three different frequencies.

Because of the two resonant frequencies f_1 and f_2, it is usually possible to identify two distinct modes of oscillation in the Klystron: one set of voltages reduces the total phase shift of the system to some multiple of $2\pi f_1$, while a second set reduces to multiples of $2\pi f_2$.

Thus, the effect of voltage appears as indicated in Fig. 9.6, showing two peaks instead of one, as might have been anticipated. The plot of $(sc/v_0\lambda)$ then has two almost parallel lines instead of one. They are practically parallel because their slope is λ/s, and although λ is slightly different in the two modes since f_1 and f_2 are different, the variation is barely distinguishable. It is of the order of 1000 cps out of 3000 mcps.

9.10 Frequency Modulation of the Klystron

Assume a Klystron that is operating at some voltage E_0 that permits oscillation. In this article, the effect of small changes in the accelerating voltage E_0 will be investigated.

The total phase shift for a Klystron oscillator was shown to be

$$\frac{2\pi sc}{\lambda v_0} + \Theta_{fb} + \frac{\pi}{2} = 2\pi N \tag{9.23}$$

This may be written in a more convenient form as

$$\frac{K}{E_0^{1/2}} + \Theta_{fb} + \frac{\pi}{2} = 2\pi N \tag{9.24}$$

where

$$K = \left(\frac{500s}{\lambda}\right)2\pi \tag{9.25}$$

and

$$\frac{K}{E_0^{1/2}} = \frac{2\pi sc}{\lambda v_0} \tag{9.26}$$

The differential of Eq. (9.24) is

$$-\frac{K}{2E_0^{3/2}}\,dE_0 + d\Theta_{fb} = 0 \tag{9.27}$$

Rearranging terms somewhat yields

$$\frac{1}{2}\left(\frac{K}{E_0^{1/2}}\right)\frac{dE_0}{E_0} = \left(\frac{d\Theta_{fb}}{df}\right)df \tag{9.28}$$

Substituting Eq. (9.26) for the bracketed quantity on the left-hand side of Eq. (9.28) yields

$$\left(\frac{\pi sc}{\lambda v_0}\right)\frac{dE_0}{E_0} = \left(\frac{d\Theta_{fb}}{df}\right)df \qquad (9.29)$$

In a series resonant circuit, near resonance, the phase angle Θ is defined by the relationship

$$\Theta = \tan^{-1}\frac{\omega L - \dfrac{1}{\omega C}}{R} \qquad (9.30)$$

This is a justifiable analogy to make with the Klystron oscillator. For small angles, the angle and tangent are approximately equal

so that $$\Theta = \tan\Theta = \frac{\omega L - (1/\omega C)}{R} \qquad (9.31)$$

or $$\Theta = \frac{2\pi fL - (1/2\pi fC)}{R} \qquad (9.32)$$

Take the derivative of Eq. (9.32) with respect to frequency, obtaining

$$\frac{d\Theta}{df} = \frac{2\pi L}{R} + \frac{1}{2\pi f^2 RC} \qquad (9.33)$$

However, at resonance, the inductive and capacitive reactances are equal, so that the equation for the derivative becomes

$$\frac{d\Theta}{df} = \left(\frac{2\pi f_r L}{R}\right)\frac{1}{f_r} + \left(\frac{2\pi f_r L}{R}\right)\frac{1}{f_r} \qquad (9.34)$$

Moreover, the quantities in brackets equal the circuit Q, that is

$$Q = \frac{2\pi f_r L}{R} \qquad (9.35)$$

Therefore, the final form of Eq. (9.33) is

$$\frac{d\Theta}{df} = 2\frac{Q}{f_r} \qquad (9.36)$$

Now substitute this relation into Eq. (9.29) for the term $d\Theta_{fb}/df$, which yields

$$\left(\frac{\pi sc}{\lambda v_0}\right)\frac{dE_0}{E_0} = \left(\frac{2Q}{f_r}\right)df \qquad (9.37)$$

Rearranging terms to obtain the change of frequency resulting from a change in accelerating voltage gives

$$\frac{df}{dE_0} = \frac{\pi s c f_r}{2\lambda v_0 Q} \tag{9.38}$$

The ease with which a Klystron may be frequency-modulated may be determined for a typical 10-cm Klystron, for which

$$s = 0.0254 \text{ m} \qquad E_0 = 1000 \text{ volts}$$
$$f_r = 3000 \text{ mcps} \qquad \lambda = 10 \text{ cm}$$
$$Q = 3000$$

Substitution of these values into Eq. (9.38) yields

$$\frac{df}{dE_0} \simeq 6000 \text{ cps/volt}$$

That is, the frequency changes 6000 cps for each volt change in accelerating voltage. Thus, it is evident that the Klystron is eminently suited for frequency modulation applications, or any place where easily varied UHF signal frequency is required. For ordinary FM transmission, a frequency deviation of 75,000 cps either side of the center frequency is needed. A Klystron would require only about a 12.5-volt signal for modulation to fully occupy the allotted bandwidth.

9.11 Tuning

The two-cavity Klystron is usually operated at maximum efficiency, over-bunching seldom occurs, and the electrical tuning range is very restricted within any given mode. One method of electrical tuning is based upon the fact (proved in Chap. 11) that a space filled with electrons has a different dielectric constant from its free-space value. This fact makes it possible to tune a Klystron by varying the beam current, which is controlled by varying the number of electrons in the beam. This method is especially applicable to tubes having high beam currents, since the effect depends upon the magnitude of the current, not the relative change in magnitude.

Cavity resonators with re-entrant shapes can be tuned mechanically by varying the capacitance between the cavity grids, or by changing the cavity volume. The latter method is generally used by tubes of glass construction, but all metal tubes may use either technique.

Klystrons generally include the resonator as an integral part of the complete tube. Flexible diaphragms are used to vary the grid spacing, thus changing the capacitive loading on the resonator, which in turn tunes the Klystron. Although the cavity volume varies somewhat as the diaphragms are moved, the change in capacitance has a much greater effect, obscuring the effect of the volume change. Thus, reducing the grid spacing reduces the resonant frequency. A typical mechanical tuning arrangement is shown in Fig. 9.9.

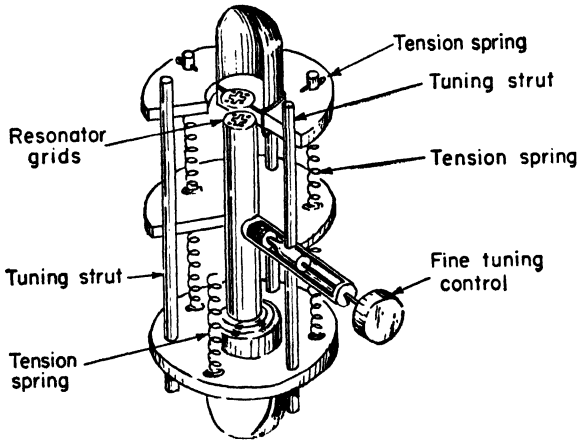


Fig. 9.9. Klystron tuner (adapted from *Klystron Technical Manual*, Sperry Gyroscope Co.).

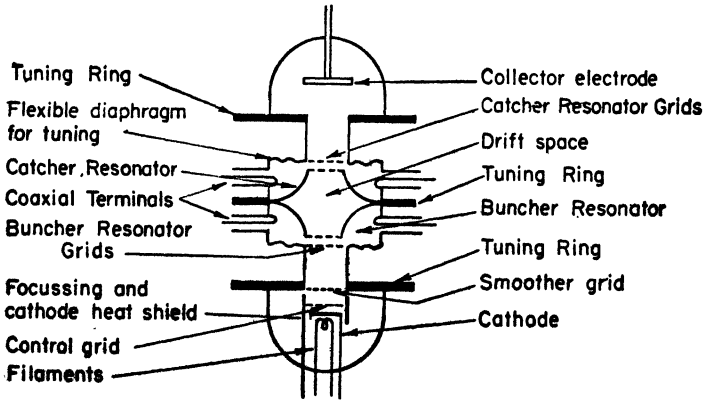


Fig. 9.10. Schematic view of a two-cavity Klystron (adapted from *Klystron Technical Manual*, Sperry Gyroscope Co.).

9.12 Mechanical Features

By way of summarizing the material covered so far, the essential constructional features of a two-cavity Klystron are shown in Fig. 9.10.

9.13 Reflex Klystron

A variation of the basic two-cavity Klystron discussed in the preceding pages is the *reflex Klystron.* In this tube one cavity is removed and a *repeller* electrode (also called a *reflector*) is added. The tube appears schematically as shown in Fig. 9.11. The single remaining cavity functions as both buncher and catcher, and although the surface distinctions appear to be rather great, the operation is closely analogous to that of the ordinary Klystron.

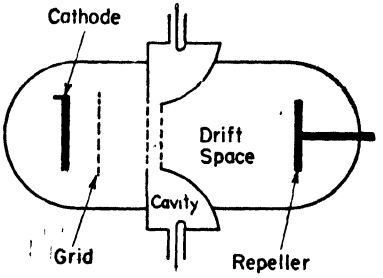


Fig. 9.11. Reflex Klystron—schematic view.

Electrons are emitted from the gun structure and accelerated by the steady field between cathode and the cavity. They are velocity-modulated by the voltage between the cavity grids and pass into the

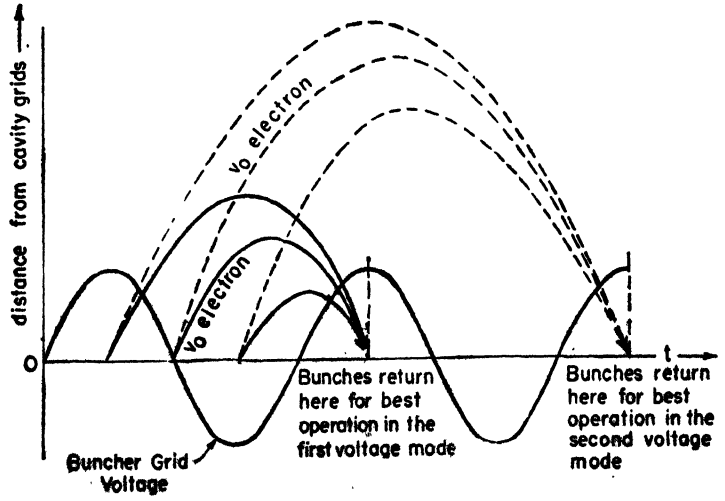


Fig. 9.12. Mechanism of bunching in a reflex Klystron.

drift space with varying velocities. As they approach the repeller, which is negatively polarized with respect to the cavity, the electrons meet a force which opposes their forward motion and causes them to stop, turn around, and return toward the cavity grids. The process of forming a bunch is illustrated in Fig. 9.12.

In the bunching process, equal numbers of electrons are decelerated and accelerated so that the cavity does not expend energy in the formation of bunches. However, the repeller is continually delivering energy to the electrons, and it is this energy, plus some acquired from the d-c field, which the electrons give up upon their return to the cavity and which is extracted as useful output.

In order for the cavity to remove energy from the bunches that return to the cavity, the time of return must coincide with a maximum retarding field between the cavity grids, as indicated in Fig. 9.12. It is evident then, that there will be several values of repeller voltage that will return the bunches at a proper time. These different operating possibilities are called *voltage modes.*

After passing back through the cavity, very few of the electrons have sufficient energy to return to the cathode. Consequently, they may fall over to the sides of the cavity and return through the source supplying the accelerating voltage, or they may return through the control grid circuit. A grid current meter can be used to indicate oscillation because, as oscillations build up, fewer electrons have sufficient energy to reach the control grid, causing the grid current to decrease.

9.14 Voltage Modes

As previously mentioned, effective operation of the reflex Klystron may be obtained whenever the bunches return to the grids, when the cavity voltage is a negative maximum. Because of the inherent 90° (¼ cycle) phase shift associated with the mechanism of bunching, the first mode occurs when the bunch returns at the end of ¾ of a cycle. The second mode occurs when the bunch returns in 1¾ cycles, and thereafter at one-cycle intervals. By varying the repeller voltage, the time of flight of the electron can be varied and the different modes of operation obtained as shown in Fig. 9.13.

The frequency at the point of maximum power output is the same for all of the modes, being the resonant frequency of the cavity. Operation in a different voltage mode does not change this frequency of resonance.

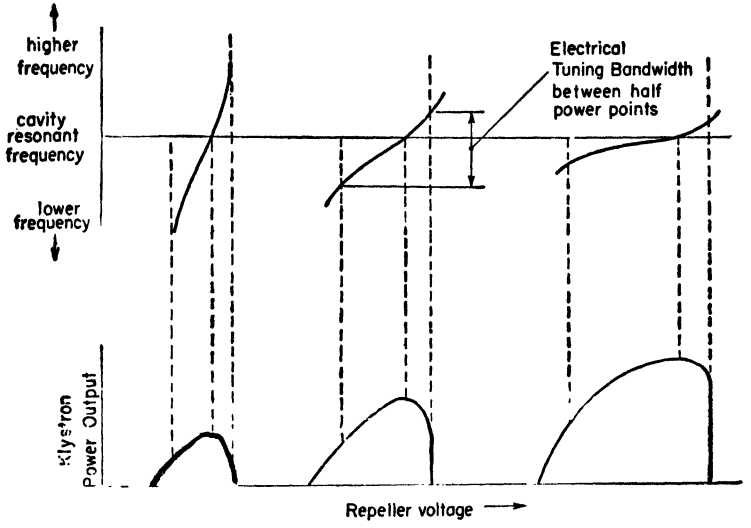


Fig. 9.13. Reflex Klystron characteristics.

Unlike the frequency, the maximum power output is different for each mode. The maximum output available is limited by two factors: losses in the cavity, and over-bunching. The latter limitation exists because, as the oscillation builds up, the cavity grid voltage (bunching voltage) likewise increases, causing the bunches to form too soon. Thus, less energy is delivered to the cavity and the growth of the oscillation is limited. In the higher modes, less bunching voltage is required because the bunches must be formed more slowly. Consequently, the oscillation is limited sooner than it would be at a lower mode. Hence, the output in the lower modes is greater than in the higher modes as was indicated in Fig. 9.13.

The humps shown in Fig. 9.13, representing oscillator power output as a function of the repeller voltage, are not symmetrical. They tend to rise slowly on the low-frequency side and fall off rapidly on the high-frequency side. The reason for this may be explained as follows. When the repeller voltage is made more negative than its value for maximum output, operation is no longer at the resonant frequency of the cavity so that the bunching voltage on the cavity grids is necessarily diminished, to an extent governed by the cavity Q. The drop in bunching voltage causes the bunches to take longer to form. However, the repeller voltage has also been decreased,

making it more negative with respect to the cavity, which causes the electrons to be returned sooner than when operating at maximum output. These two effects combine to make the output drop off very rapidly. On the other hand, reducing the voltage drop between repeller and cavity by making the repeller less negative causes the electrons to be returned later than they would be for maximum power output However, since operation is no longer at the resonant frequency of the cavity, the bunching voltage is reduced and tends to make the bunches form at a later time. These two effects serve to partially counterbalance one another so that the point of maximum bunching tends to remain between the cavity grids. Consequently, the output drops off relatively gradually on this side of the mode.

9.15 Electrical Tuning of the Reflex Klystron

A limited amount of electrical tuning may be obtained within each of the voltage modes by "pulling" the Klystron. That is, by returning the bunches a little earlier or later than normal, the frequency may be pulled slightly off the resonant frequency of the cavity. The extent to which the Klystron may be pulled depends largely upon the cavity Q, since that governs the magnitude of the bunching voltage at frequencies off resonance. In any case, the shift is small compared to the operating frequency.

The time required for bunch formation can be varied by either of two voltages, the accelerating voltage (between cavity and cathode) or the repeller voltage (between repeller and cavity). Variations in the accelerating voltage produce two effects which *tend* to cancel one another. An increase in accelerating voltage gives the electrons a higher velocity, which would, ordinarily, make it more difficult for them to be reflected by the repeller, thus increasing their transit time. Unfortunately, the increase in cavity voltage also increases the repeller voltage negatively, causing it to exert more force on the electrons and return them sooner. The two effects are subtractive, each tending to nullify the other.

However, varying the repeller voltage produces only the second effect, making this form of electrical tuning much easier than by changing the d-c cavity voltage. The conveniently high-frequency sensitivity resulting from variations in repeller voltage (of the order of ½ mcps per volt) is offset somewhat by the necessity for a very well-regulated power supply.

Assuming the tuning bandwidth to be the band of frequencies between half-power points for that mode, it is apparent from inspection of Fig. 9.13 that tuning over a wider range is possible in the higher modes than in the lower modes. This results from the mechanism of bunching. In the higher modes, the bunches form and then de-bunch more slowly than in the lower modes. Consequently, they can be returned at a slightly different time in the cycle and remain comparatively well bunched, keeping the power output high. Rapid de-bunching in the lower modes causes the output to drop off rapidly if the bunch is returned slightly before or after normal.

Because the electrical tuning bandwidth is greatest for the weakest (higher) modes, a compromise between power output and tuning range is often necessary.

9.16 Summary of Klystron Applications

The discussion so far has been largely concerned with the Klystron as an oscillator or power amplifier. However, it can be used very effectively as a low-power amplifier. The difference in operation lies principally in the degree of velocity modulation used. For oscillators and power amplifiers the bunching voltage is fairly large so that the electron beam is intensity-modulated to the fullest extent. This produces pulses of catcher current.

If a very small signal is applied to the buncher grids, the intensity modulation is slight and very nearly a replica of the input. Under this condition, the Klystron can be used as a small signal amplifier with a voltage amplification of about 20 at frequencies as high as 3000 mcps.

The greater the degree of velocity modulation, the more distorted the catcher current becomes, and, as previously pointed out, the more suitable the Klystron becomes as a frequency multiplier. When operated as an oscillator or amplifier, both cavities are tuned to the same frequency, but when operated as a frequency multiplier, the catcher is tuned to a higher harmonic.

By appropriate connections, the Klystron may also be operated as a detector. Furthermore, while linear amplitude modulation is difficult, frequency modulation is readily produced.

Thus, the Klystron finds application as an oscillator, amplifier, detector, and modulator.

9.17 Other Velocity-modulated Tubes

McNally tube. Very similar to the reflex Klystron, but part of the cavity is external and removable from the evacuated part of the tube. Part glass, part metal construction.

Shepherd-Pierce tube. All metal construction, but otherwise similar to the reflex Klystron and McNally tube. The tuning device is an integral part of the tube in the form of a bowed strut that changes the grid spacing. Fits a standard octal socket.

Haeff UHF tube. Uses a cavity for energy removal from an electron beam, but the electron beam is controlled by a single grid instead of a second cavity. Thus, it requires a fair amount of driving power. It resembles a Klystron with the buncher cavity replaced by a control grid.

Hahn-Metcalf tube. Uses velocity modulation, but otherwise quite dissimilar from the Klystron. For a brief discussion refer to *UHF Techniques* by Brainerd, *et al* (Van Nostrand), or to the basic reference listed at the end of this chapter.

REFERENCES

Condon, E. U., "Electronic Generation of Electromagnetic Oscillations, *J. Applied Phys.*, **11**, 502 (1940).

Haeff, A. V., "An Ultra High Frequency Power Amplifier of Novel Design," *Electronics*, **12**, 30 (1939).

Hahn, W. C., "Small Signal Theory of Velocity Modulated Electron Beams," *G.E. Review*, **42**, 258 (1939).

Hahn, W. C., and Metcalf, G. F., "Velocity Modulated Tubes," *Proc. IRE*, **27**, 106 (1939).

Harrison, A. E., *Klystron Technical Manual*, Sperry Gyroscope Co.

Varian, R. H., and Varian, S. F., "A High Frequency Amplifier and Oscillator," *J. Applied Phys.*, **10**, 140, 321 (1939).

Webster, D. L., "Cathode Ray Bunching," *J. Applied Phys.*, **10**, 501 (1939).
"The Theory of Klystron Generators," *J. Applied Phys.*, **10**, 864 (1939).

A very complete discussion of reflex Klystron tube types will be found in *Bell System Tech. J.*, July (1947).

CHAPTER 10

MAGNETRON OSCILLATORS

A DISTINCT transmitter problem exists in the microwave region due to the failure of ordinary triode oscillators to produce sufficient power at the required frequencies. Since reduction in tube size to increase the upper frequency limit inhibits the allowable power dissipation, it becomes necessary to resort to transit time oscillators of the Klystron or magnetron type.

The Klystron, or velocity-modulated tube, has found a wide application as a local oscillator in UHF and microwave receivers. In the few years following its development, it was not particularly adapted to pulsed operation and it did not have a very large power output. As a consequence, it was not suitable for general transmitter use. Subsequent development has produced Klystrons having 18 megawatts of peak power in pulsed operation, and the earlier limitations have been effectively removed. Furthermore, recent advances in the design of microwave *lens* systems, which are being extensively used in microwave relay stations, give tremendous power gains, because of increased directivity, and allow low-power Klystrons (as low as 1 watt) to be used effectively as transmitters.

The need for additional high-power microwave transmitting tubes was met with the development of the magnetron. It is capable of delivering exceptionally large peak powers, at comparatively high efficiencies. As a result, it is used almost exclusively in high-power microwave transmitters. The magnetron is not readily tunable and is subject to considerable noise. Therefore, it does not encroach upon the Klystron in local oscillator applications.

10.1 The Magnetron

In its most fundamental form, the magnetron is a diode of cylindrical construction. The voltage between the anode and cathode (which is coaxial with the anode) produces an electric field that is

radial. A uniform magnetic field is applied within the anode-cathode space, longitudinally, that is, parallel to the cathode. The general configuration of the prototype magnetron appears as shown in Fig. 10.1. The relative orientations of the electric and magnetic field intensities are shown in Fig. 10.2. In general, for use as an oscillator, the magnetic field is maintained constant while the electric field has an alternating *RF* field superimposed on the d-c value.

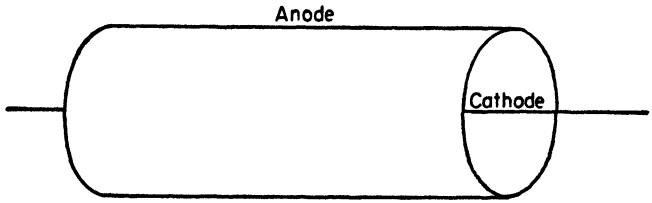


Fig. 10.1. Prototype magnetron structure.

Magnetrons may operate with a single anode, as shown in Fig. 10.1, but more generally, the anode is split into segments. It is then called a *split-anode magnetron.* At comparatively low frequencies, the tuned circuit for the oscillator may be provided by resonant lengths of transmission line (Lecher wires). In the microwave region the tuned circuit is conventionally a resonant cavity.

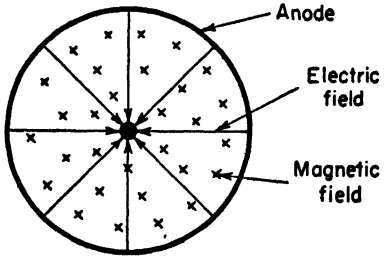


Fig. 10.2. End view of magnetron showing the radial electric field and the axial magnetic field.

In some cases magnetrons are of all metal construction, but in others, notably for CW magnetrons at the lower frequencies, the magnetron may be enclosed within a glass envelope. The magnetron is used almost exclusively as a transit time oscillator, although it may be made to oscillate in the conventional negative resistance manner. In any case, a study of the electronic trajectories in the combined electric and magnetic field is necessary in order to understand the basic mechanism of the energy transformation from d-c to *RF*. A detailed mathematical analysis will be carried through in the next few articles.

10.2 Equations of Motion in Combined Fields

Although the magnetron is almost universally constructed in the form of a cylindrical diode, theoretical consideration of a magnetron composed of two parallel, plane electrodes is just as instructive and a great deal less involved from the mathematical standpoint. Mathematics in cylindrical coordinates are, in this case, just too much bother for the net worth obtained.

Several assumptions are necessary, their function principally being to orient the coordinate system in such a way that the mathematical processes are not too cumbersome. Assume a coordinate system as shown in Fig. 10.3 so that the electric field is parallel to the Z axis while the magnetic field and cathode are along the Y axis. Both electrodes are assumed to be plane, parallel, and of infinite extent.

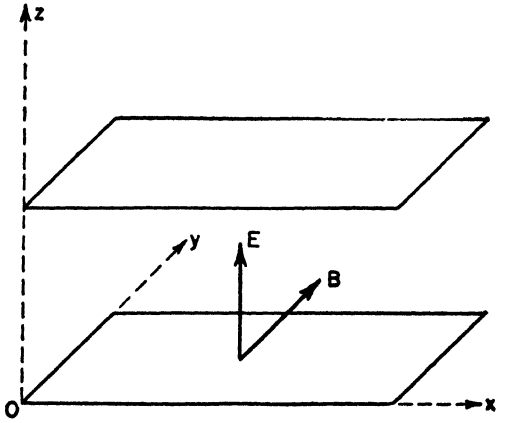


Fig. 10.3. Coordinate system assumed for the parallel-plane-electrode magnetron.

An electron which is emitted by the cathode enters the combined field and is acted upon by two separate external forces, one due to each field. The electron is assumed to have a mass m and a charge e. The electric and magnetic field intensities are denoted by $\mathbf{E}$ and $\mathbf{H}$, respectively.

The electron moving in the electric field is acted upon by a force, independent of the electron velocity, of strength $e\mathbf{E}$ and directed opposite to the direction of the electric field intensity vector. If

the field is constant and uniform, the force has the same general character as gravitational attraction.

The force exerted by the magnetic field is somewhat different. It depends upon the magnetic field strength, charge, and the relative velocity between the field and charge according to the general equation

$$\mathbf{f} = e(\mathbf{v} \times \mathbf{B}) \tag{10.1}$$

For a homogeneous isotropic medium, the permeability may be brought outside the bracketed quantity, giving

$$\mathbf{f} = \mu_0 e(\mathbf{v} \times \mathbf{H}) \tag{10.2}$$

Thus, the total force equation for the electron when it is emitted into the combined field is

$$\mathbf{f} = e\mathbf{E} + \mu_0 e(\mathbf{v} \times \mathbf{H}) \tag{10.3}$$

Carry out the indicated vector cross product and equate like components, as follows:

$$f_x = eE_x + \mu_0 e(v_y H_z - v_z H_y) = m\frac{dv_x}{dt} \tag{10.4}$$

$$f_y = eE_y + \mu_0 e(v_z H_x - v_x H_z) = m\frac{dv_y}{dt} \tag{10.5}$$

$$f_z = eE_z + \mu_0 e(v_x H_y - v_y H_x) = m\frac{dv_z}{dt} \tag{10.6}$$

However, it will be recalled that the coordinate axes were picked in such a way that the electric field had only a Z component while the magnetic field had only a Y component. Consequently, the equations of motion above reduce to

$$f_x = -\mu_0 e v_z H_y = m\frac{dv_x}{dt} \tag{10.7}$$

$$f_y = 0 = m\frac{dv_y}{dt} \tag{10.8}$$

$$f_z = eE_z + \mu_0 e H_y v_x = m\frac{dv_z}{dt} \tag{10.9}$$

They may be put into a more convenient form by dividing through by the mass of the electron and letting

$$\frac{e}{m}\mu_0 H = \omega \qquad \text{where } H = H_y \tag{10.10}$$

$$\frac{e}{m}E = a \qquad \text{where } E = E_z \tag{10.11}$$

The equations of motion then become

$$\frac{dv_x}{dt} = -\omega v_z \tag{10.12}$$

$$\frac{dv_y}{dt} = 0 \tag{10.13}$$

$$\frac{dv_z}{dt} = a + \omega v_x \tag{10.14}$$

10.3 Solution of the Equations of Motion

Integrate the equation for the acceleration in the Y direction, Eq. (10.13). According to (10.13),

$$a_y = \frac{dv_y}{dt} = 0$$

which, when integrated, yields

$$v_y = K = \text{constant} \tag{10.15}$$

Differentiate Eq. (10.14) with respect to time, obtaining

$$\frac{d^2v_z}{dt^2} = \omega \frac{dv_x}{dt} \qquad \text{since } a = \text{constant} \tag{10.16}$$

However, according to Eq. (10.12)

$$\frac{dv_x}{dt} = -\omega v_z$$

Consequently, Eq. (10.16) becomes

$$\frac{d^2v_z}{dt^2} = -\omega^2 v_z$$

or, rearranging terms,

$$\frac{d^2v_z}{dt^2} + \omega^2 v_z = 0 \tag{10.17}$$

This is a second order linear differential equation with constant coefficients with a solution that can be obtained by any standard technique, of the form

$$v_z = A \sin \omega t + B \cos \omega t \tag{10.18}$$

Complete evaluation of the constants A and B depends upon the specification of the initial conditions of the velocity components. It

is customary, in order to simplify the mathematical procedure, to assume that the electrons are emitted from the cathode with zero velocity. Taking the moment of emission as the initial instant, then the problem may be nailed down to the specific case where,

$$\text{when } t = 0, \qquad \text{then } v_x = v_y = v_z = 0$$

Substituting this initial condition into Eq. (10.15) for the velocity in the Y direction yields

$$K = 0, \qquad \text{or} \qquad v_y = 0 \text{ at all times} \tag{10.19}$$

Substituting the initial condition into the equation for the Z component of the velocity gives

$$0 = 0 + B \qquad \text{or} \qquad B = 0$$

Consequently, the expression for v_z is

$$v_z = A \sin \omega t \tag{10.20}$$

Differentiate this equation, obtaining

$$\frac{dv_z}{dt} = A\omega \cos \omega t$$

But, according to Eq. (10.14)

$$\frac{dv_z}{dt} = a + \omega v_x$$

Hence, equating the two expressions for the acceleration in the Z direction yields

$$a + \omega v_x = A\omega \cos \omega t$$

Substitution of the initial condition into this equation gives

$$A = \frac{a}{\omega}$$

Then, substituting for A in Eq. (10.20) yields

$$v_z = \frac{a}{\omega} \sin \omega t \tag{10.21}$$

Differentiating with respect to time, as before,

$$\frac{dv_z}{dt} = a \cos \omega t = a + \omega v_x$$

This equation may be solved for v_x, producing

$$v_x = \frac{a}{\omega}(\cos \omega t - 1) \tag{10.22}$$

In summary, then, the velocity equations are

$$v_x = \frac{a}{\omega}(\cos \omega t - 1)$$

$$v_y = 0$$

$$v_z = \frac{a}{\omega} \sin \omega t$$

Integrate all three of these equations to obtain the expressions for the electron displacement. Since the center of coordinates corresponds to the point of emission of the electron from the cathode, then it follows that at $t = 0$, $x = y = z = 0$. Hence,

$$y = 0 \tag{10.23}$$

$$x = -\frac{a}{\omega^2}(\omega t - \sin \omega t) \tag{10.24}$$

$$z = \frac{a}{\omega^2}(1 - \cos \omega t) \tag{10.25}$$

Equations (10.24) and (10.25) are the parametric equations for a cycloid in the X-Z plane, with the general progession along the X axis. The radius of the cycloid is the coefficient

$$P_c = \frac{a}{\omega^2}$$

but $$a = \left(\frac{e}{m}\right)E \quad \text{and} \quad \omega = \left(\frac{e}{m}\right)\mu_0 H$$

so that $$P_c = \left(\frac{m}{e}\right)\left(\frac{1}{\mu_0{}^2}\right)\left(\frac{E}{H^2}\right)$$

10.4 Deductions from the Solution

When the parametric equations for the cycloid are plotted, the path of the electron, relative to the coordinate axes, is as shown in Fig. 10.4, where

$$a = \left(\frac{e}{m}\right)E$$

$$\omega = \left(\frac{e}{m}\right)\mu_0 H$$

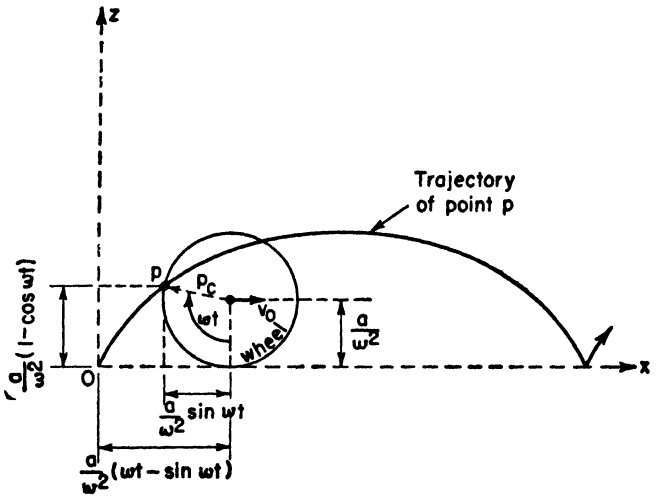


Fig. 10.4. Cycloidal path of electron in the parallel-plane-electrode magnetron, with respect to the coordinate system.

From the cycloidal path it is seen that the radius of the cycloid, as previously calculated, is

$$P_c = \left(\frac{m}{e}\right)\left(\frac{E}{B^2}\right) \qquad \text{where } B = \mu_0 H \tag{10.26}$$

and the angular velocity of the electron is

$$\omega = \left(\frac{e}{m}\right)\mu_0 H \tag{10.27}$$

There are several important deductions to be made. From Eqs. (10.26) and (10.27) it is apparent that the relative dimensions of the cycloid are determined wholly by the electric and magnetic field intensities. Since a cycloid is generated by a point on a rolling wheel, as indicated in Fig. 10.4, it can be deduced that, regardless of the location of the point on the wheel, the total *horizontal* distance of travel (that is, along the X axis) is the same for all points on the wheel. It would simply be the total distance that the center of the wheel moves in a given length of time. Figure 10.5 shows several typical electron orbits.

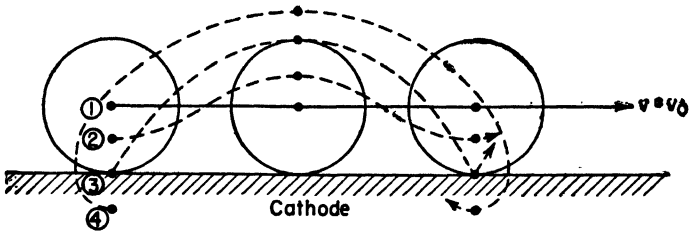


Fig. 10.5. Method of generating typical electron orbits.

The period of the rolling wheel (T_0) is

$$T_0 = \frac{1}{\text{frequency}} = \frac{2\pi}{\text{angular velocity}} = \frac{2\pi}{\omega} \tag{10.28}$$

hence
$$T_0 = \left(\frac{m}{e}\right)\left(\frac{2\pi}{\mu_0}\right)\frac{1}{H} \tag{10.29}$$

Thus, the period is determined solely by the magnetic field; the electric field has no effect. Actually, it is the period that would be obtained if the electron were acted upon by the magnetic field alone.

The velocity of the center of the wheel, v_0, is readily calculated now since

$$v_0 = \frac{\text{wheel circumference}}{\text{period}} = \frac{2\pi P_c}{T_0} \tag{10.30}$$

$$v_0 = \left(\frac{e}{m}\right)^2\left(\frac{E}{B}\right) \qquad \text{where } B = \mu_0 H \tag{10.31}$$

Using this as a reference velocity, Fig. 10.6 illustrates the effect of having electrons with initial horizontal velocities that are greater or less than this base value.

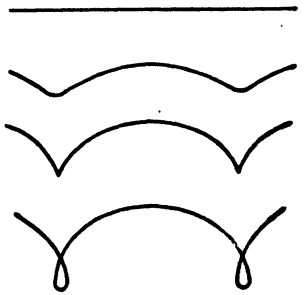


Fig. 10.6. Orbital configurations for different initial velocities of the electrons.

The electron orbit, with respect to the electrodes and fields, is shown in Fig. 10.7. In the case of cylindrical construction, the electric field is radial and the magnetic field is axial. Neglecting space charge effects, as was done in the parallel plane case, the electron orbit approximates an epicycloid generated by rolling a wheel around the cathode. The orbit is not exactly epicycloidal because the radial motion is not simple

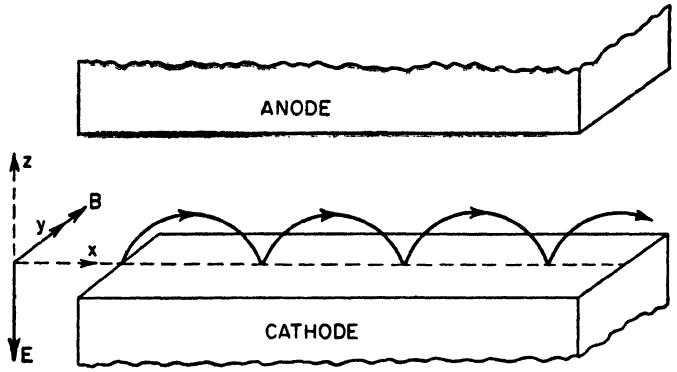


Fig. 10.7. Orientation of the electron trajectory in a d-c magnetron of parallel-plane-electrode construction.

harmonic motion. This is because the d-c electric field varies with the radius. However, the approximation is pretty good when the cathode-to-anode radius is large. A typical electron orbit in a cylindrical magnetron is shown in Fig. 10.8.

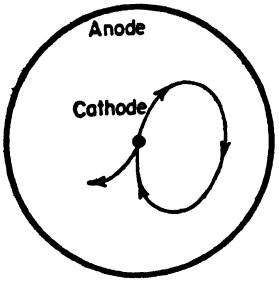


Fig. 10.8. Typical electron orbit in a cylindrical d-c magnetron.

10.5 Cutoff

From the simplified treatment given so far it is apparent that, for a given electric field intensity E, the magnetic field can be adjusted to such a value that it causes the electrons to miss the anode completely and return to the cathode. When this happens, the plate current ceases and the magnetron is cut off. Consequently, the value of the magnetic field intensity which separates conduction from nonconduction is called the critical magnetic field and denoted by H_c, or B_c in terms of the flux density. Several different electron orbits for different values of magnetic flux density above and below B_c are shown in Fig. 10.9.

Hypothetically, the curve of plate current as a function of magnetic field intensity appears as shown by the dotted line of Fig. 10.10. Practically, the square cutoff characteristic is rounded off as shown by the solid line curve in Fig. 10.10. The rounding results from the

fact that the assumptions made in the analysis are not exactly met. For example, space charge has been neglected and it is a comparatively important factor. Also, the mechanical problem of manufacturing a perfectly symmetrical tube with a perfectly axial filament and magnetic field is never completely met. Furthermore, the electric field will fringe somewhat at the ends of the anode structure, causing nonuniformity of the electric field.

The magnitude of the theoretical

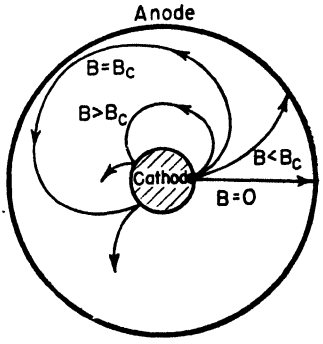


Fig. 10.9. Variation in electron orbit with magnetic flux density—illustrating the cutoff characteristic.

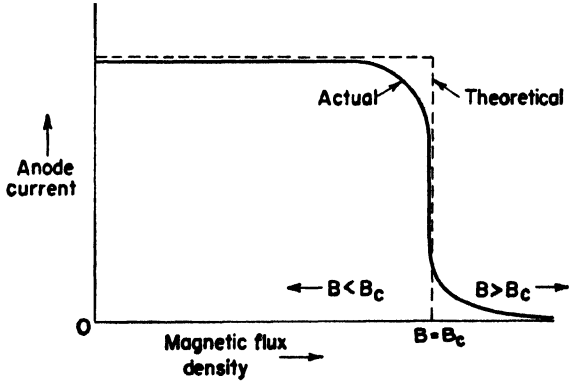


Fig. 10.10. Magnetron cutoff characteristic.

critical value of magnetic flux density may be calculated by equating the diameter of the rolling wheel to the distance between electrodes. That is,

$$S = 2P_c = 2\left(\frac{e}{m}\right)\left(\frac{E}{B^2}\right)$$

80

$$B_c = \mu_0 H_c = \sqrt{\frac{2}{S}\left(\frac{e}{m}\right)E}$$

10.6 Principal Modes of Oscillation

The d-c magnetron, which has occupied our attention in the preceding articles, may be converted into an *RF* oscillator by introducing *RF* fields into the cathode-anode space. This may be done by applying an *RF* voltage between anode and cathode from a resonant circuit. Another method of introducing this field is to split the anode into segments and connect a resonant circuit between these anode segments. In any case, the production of oscillations depends upon the manner in which the interaction of the electron motion with the alternating field takes place.

There are three distinct possible modes of oscillation in a magnetron that have been identified. They are generally listed as follows:

(1) Dynatron (negative-resistance) magnetron.
(2) Cyclotron frequency magnetron.
(3) Traveling-wave magnetron.

Operation in the first and last modes can occur *only* in split-anode magnetrons, whereas the second mode can be produced in either a single- or a split-anode magnetron. The traveling-wave magnetron is the type most generally, almost exclusively, used commercially due to its much higher efficiency. Both the traveling-wave and cyclotron modes are frequently lumped together in the classification of "electron resonance." Although the term is not exactly a misnomer, since both modes are transit time phenomena, the expression frequently leads to misconceptions that can be avoided by using the classification given here, which is the terminology employed by the Bell Telephone Laboratories and that which will probably receive the most widespread acceptance.

10.7 Dynatron Operation of Magnetrons

The operation of a magnetron in the dynatron mode hinges upon the development of a negative resistance characteristic between the two segments of a split-anode magnetron. An ordinary tuned circuit connected between these segments then produces an oscillator. A typical connection is shown in Fig. 10.11. Because of the similarity of the circuit configuration to the usual push-pull oscillator, it is frequently called the *push-pull connection.*

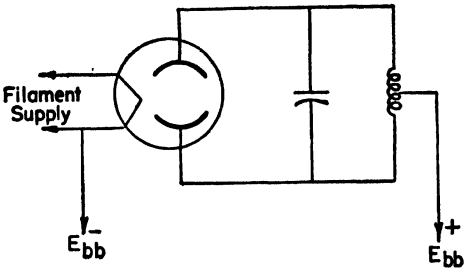


Fig. 10.11. Schematic connections for dynatron operation of a magnetron.

The efficiency and power output in this mode are limited by the same factors that apply to ordinary negative grid oscillators. As a matter of fact, the requirements for the Q of the tuned circuit are even more severe than for the usual triode oscillator. Oscillations have been produced all the way from the audio region up to as high as 1000 mcps. However, at 600 mcps, a power output of 100 watts at 25 per cent efficiency is about the best that can be obtained practically. Consequently, it offers no especial advantage over the ordinary triode oscillator, and suffers from several disadvantages, notably the necessity for a strong magnetic field, which do not apply to triodes. Nevertheless, the possibility of future development warrants its inclusion here.

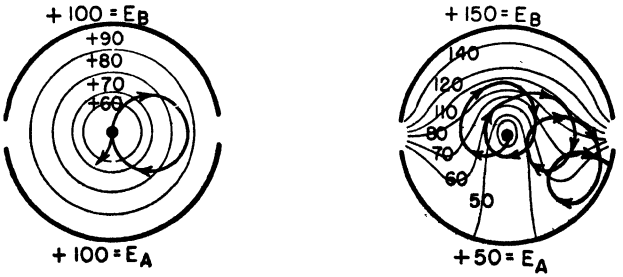


Fig. 10.12. (a) Potential lines when both anodes are at the same potential. Electron orbit is as shown. (b) Equipotential lines when anode segments are at different potentials. Note that the electrons go to the anode that is at the lowest potential.

The reason for the negative resistance effect may be determined from the diagrams in Fig. 10.12. This figure shows typical hypothetical electron trajectories for the case where the magnetic field is considerably larger than the critical value B_c. When the anode segments are at equal voltages with respect to the cathode, the electrons travel in symmetrical paths as shown in Fig. 10.12(a) and never reach the anode. However, when the voltage of one segment is increased, and the other is decreased, the field distribution is altered as shown by Fig. 10.12(b). The electric field is such that the forces on the electron cause it to strike the anode segment that is at the *lowest potential.* Consequently, reducing the voltage on one segment increases its current, and conversely, over a restricted range of voltages.

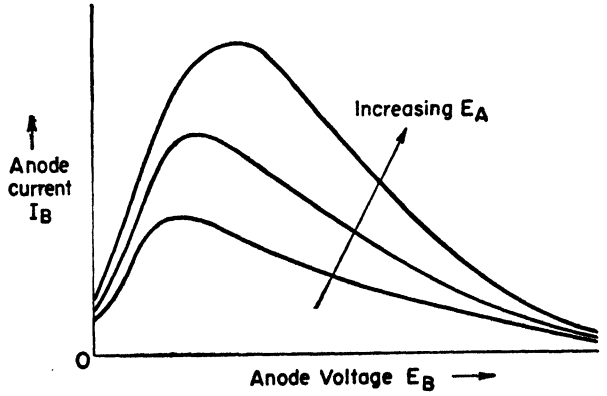


Fig. 10.13. Plate-current, plate-voltage characteristic for two-anode magnetron.

A typical plot of anode current against anode voltage, for a two-anode magnetron, is shown in Fig. 10.13. Plotting the sum and the difference of the anode segment currents against the voltage between segments yields the characteristic curves of Fig. 10.14. The difference curve exhibits a region of negative resistance.

10.8 Cyclotron Frequency Oscillations

The practical significance of the magnetron arises from the fact that it effectively uses transit time to its advantage. Consequently, since the dynatron mode is limited to a large extent by transit time,

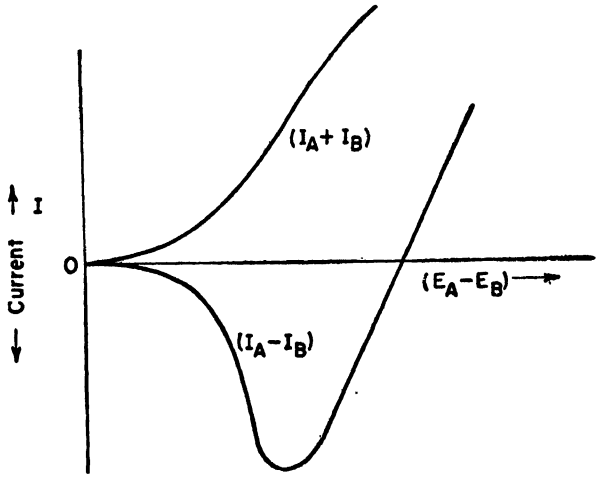


Fig. 10.14. Negative resistance in the static characteristics of a two-anode magnetron.

just as in the ordinary triode oscillator, it is not the most important mode of oscillation.

The first of the transit time modes in the magnetron is the cyclotron frequency mode. The analysis of the mechanism of energy conversion from d-c to *RF* will be carried through for the case of parallel plane electrodes. As previously pointed out, plane-electrode magnetrons are never used practically, but the analysis is simpler and serves as a logical introduction to cylindrical magnetrons.

Assume a parallel-plane-electrode magnetron in which the flux density is adjusted to a point that will insure plate current cutoff. Therefore, the electron orbit will appear as shown in Fig. 10.15. Assume that the time required for the electron to make one round trip from the cathode to the anode and return to the cathode is t_0. Then, finally, assume that an *RF* sinusoidal voltage of period t_0 is superimposed on the steady field already present.

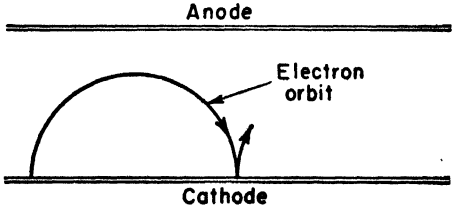


Fig. 10.15.

If an electron leaves the cathode at the time that the *RF* voltage is going through zero in the *positive direction,* the total electric field acting on the electron is greater than the d-c field. This causes the electron to go closer to the anode than it would have in the presence of the steady field alone. This is because the radius is given by

$$P_c = K\frac{E}{H^2} \qquad \text{where } K = \text{constant}$$

and is directly proportional to the electric field intensity. As a result, the electron may strike the anode, or it may miss it and return to the cathode, doing so as the *RF* field is decreasing *below* the steady value. Thus, the *RF* electric field is continuously doing work on the electron, adding energy above and beyond that imparted by the d-c field. Hence, the electron extracts energy from the source of *RF* power and gives it up by impact to the anode or the cathode. Such an electron is undesirable, naturally, and is called an *energy-removing* or *unfavorable* electron.

However, if an electron leaves the cathode a half period later than the most unfavorable electron just discussed, when the plate voltage is going through zero in the *negative* direction, then the field acting on the electron on the outgoing part of the trip is less than the steady field. Consequently, it would not go as close to the anode as it would have in the presence of the steady field alone. The field increases in the positive direction as the electron starts its return trip. Thus the electron continuously works against the *RF* field and delivers some of the energy that it received from the d-c field to the source of *RF* energy. Consequently, it is an *energy-delivering* or most *favorable* electron. Because of this loss in kinetic energy, the electron comes to rest at a point in the cathode-anode space before reaching the cathode.

As a general conclusion it may be stated that

(1) Electrons leaving the cathode when the *RF* component of the electric field is increasing in the positive direction take energy away from the source of *RF* power.
(2) Electrons leaving the cathode when the *RF* component of anode voltage is going through zero in the negative direction deliver energy to the *RF* source.

There are other important observations. The energy-removing electrons may do either of two things:

(1) They may strike the anode on their first outward trip.
(2) They may return to the cathode.

In any case, they cannot make more than one round trip. The energy-delivering electrons, after initially coming to rest at some point in the cathode-anode space, can continue to oscillate with decreasing amplitude. In each successive cycle of the electronic oscillation, the distance traveled and the transit time are diminished. This causes the electron oscillation to advance in phase relative to the *RF* voltage and it may eventually become so far advanced that the electron removes energy rather than delivering it.

In substance then, the successful operation of the magnetron oscillator in this mode depends upon the fact that the energy-delivering electrons make several excursions into the cathode-anode area and thus give up more energy than the energy-removing electrons extract in their single excursion. Statistically, the electrons are equally divided in number between the two categories.

It is designated as the *cyclotron mode* because the *RF* period and electronic period are equal. This makes adjustment of the magnetic field and tuning of the cavity a very delicate and precise job.

10.9 Factors Affecting Efficiency

It was previously shown that the radius of the cycloidal hops was given the equation.

$$P_c = \left(\frac{e}{m}\right)\frac{E}{B^2}$$

Because the favorable electrons deliver energy to the *RF* field, they come to rest at an earlier time than they would have in the absence of the alternating field. Consequently, the electronic oscillation advances in phase with respect to the *RF* voltage. Hence, during the succeeding hop, the electron is subjected to less accelerating field and more decelerating field so that the average value of the electric field intensity with which the electron interacts is less than on the preceding hop. This causes the radius of each successive hop to decrease as shown in Fig. 10.16. Thus, the reduction in amplitude of each hop follows as a natural consequence of the phase advancement.

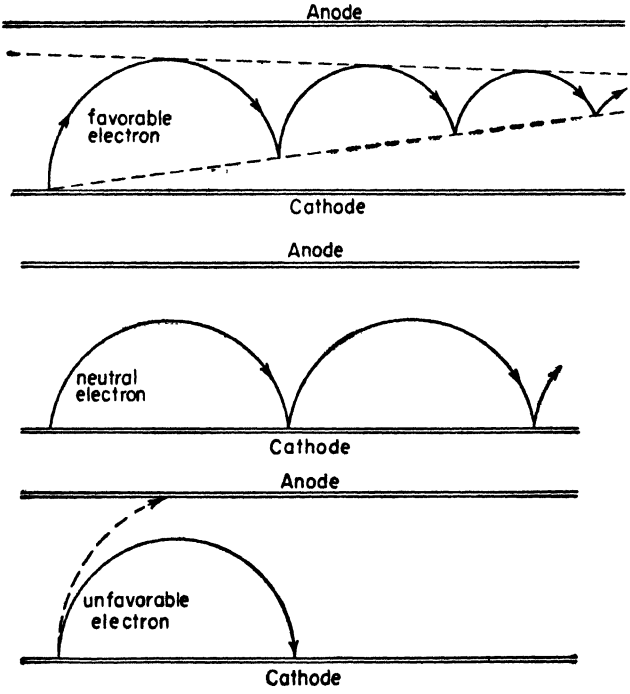


Fig. 10.16. Typical electron orbits in a parallel-plane-electrode magnetron operated in the cyclotron frequency mode.

As a result of this action, two undesirable effects can be produced. The electron can come to rest in the cathode-anode space when the radius of the cycloid becomes vanishingly small. It would then start oscillating in reverse phase, with increasing amplitude, as it picks up energy from the *RF* source. Even if this did not occur, the advance in phase would generally be so large that the originally favorable electrons would become unfavorable ones, being removed from the cathode-anode space at the expense of the oscillating field.

In order to extract a reasonable amount of power from the tube, these electrons which are becoming unfavorable must be removed from the interelectrode space before they start absorbing energy because of their precession. Sideward motion of the electrons in the parallel-plane-electrode magnetron would carry them out of the cathode-anode space in sufficient time to prevent the phase shift from becoming objectionably large.

However, the cylindrical magnetron poses a rather neat design problem. One method of solving it is to tilt the magnetic field in such a way that the electrons are given a velocity component parallel to the filament. The angle of tilt is quite critical, since the designer is faced with two mutually exclusive conditions:

(1) To remove the electron before it starts absorbing appreciable energy.
(2) To leave the electron in the interelectrode space long enough for it to give up the greatest amount of power.

Some sort of collecting plates, or *end plates*, are required at the ends of the magnetron in order to remove the electrons safely. This is necessary since the energy of the electrons striking the glass walls of the tube could be sufficiently large to melt the glass. The potential of the end plates must be below that of the anodes in order to function properly. In some cases, end plates are used without tilting the magnetic field. In such applications, it has been found that the best end plate voltage is proportional to the anode voltage.

Another undesirable characteristic of the cyclotron mode of operation arises from the necessity of employing large cycloidal hops of the electrons in order to meet the frequency requirements. The large hop produces a high kinetic energy at the moment that the electron strikes the anode, producing large heat losses and consequent lowering of efficiency.

All of these factors conspire to give a rather low operating efficiency so that this mode is *not* extensively used in current commercial magnetrons. Efficiencies of 10 to 15 per cent have been obtained at wavelengths as low as 0.64 cm.

10.10 Introduction to the Traveling-wave Magnetron

Both the cyclotron frequency mode and the traveling-wave mode of oscillation in a magnetron depend upon electron transit time. The similarity ceases at that point. The traveling-wave mode of oscillation is fundamentally different from either of the preceding two.

This mode will function *only* with split-anode construction and the conversion of energy from d-c into *RF* depends upon the interaction of the electron with the field in the air gap between anode segments. For the parallel-plane-electrode case, the field configurations appear as shown in Fig. 10.17. This *interaction* field may be resolved into

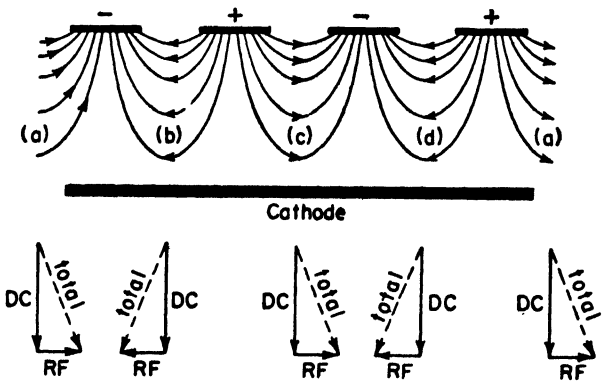


Fig. 10.17. Field configurations in a split-anode magnetron interaction field are shown in the top figure. The resultant electric field between anode segments is shown at the bottom.

components, one set parallel to the original d-c electric field (vertical field), and one set parallel to the axis of the tube (horizontal field). The dimensions and characteristics of the cycloid traced out by the electron in its trajectory are determined by the vertical electric field, assuming a constant magnetic field. The vertical electric field consists of a d-c value with a superimposed *RF* component. Over any cycle of the *RF* voltage, the average value of the total vertical field is simply the steady field. Consequently, on the average, the radius of the cycloid is unchanged throughout the trajectory.

Conversion of d-c energy into *RF* by the electron is effected by the electron's doing work against the *horizontal* component of the interaction field. Thus, it is seen that the functions of producing the cycloidal trajectory and of interchanging energy have been separated in this mode. In the cyclotron mode of oscillation, both operations occurred simultaneously with the same electric field component.

Because of the manner of energy removal from the electron, the requirement for the most favorable electron is simply that it pass through the interaction field with maximum horizontal velocity at a time when the horizontal component of the field is exerting maximum opposition. In other words, for maximum efficiency, the electron should be at the top of its hop and in the center of the air gap at the time that the interaction field has its maximum value. Consequently, there is *no* specification of the number of cycloidal hops between air

gap crossings, only that each crossing take place at a time when the interaction field has a horizontal component producing maximum opposition. Any number of cycloidal hops may intervene. Contrast this with the rigidity of the requirements placed upon the electron in the cyclotron mode of oscillation.

Small cycloidal hops are conducive to high efficiency since they reduce the kinetic energy of impact on the anode segments. Consequently, since the number and size of the hops is immaterial as far as the requirement for oscillation is concerned, small hops are generally used and an electron will make a great many small jumps in traversing one interaction field. If the horizontal component of the field is a retarding field, the electron delivers energy to the *RF* source and gradually progresses up from the cathode to the anode as shown in Fig. 10.18. Electrons in regions (a) and (c) of Fig. 10.17 would

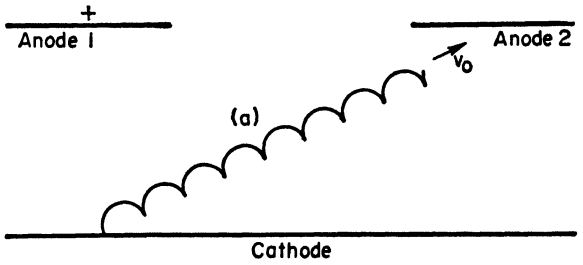


Fig. 10.18. Path of a favorable electron, in a region such as (a) or (c) of Fig. 10.17, for operation in the traveling-wave mode.

have this sort of trajectory. The drift of the electron from cathode to anode indicates the conversion of energy from d-c to *RF*. On the other hand, electrons in regions (b) and (d) are acted upon by accelerating forces due to the horizontal component of the field, extract energy from the *RF* source, and progress as shown in Fig. 10.19.

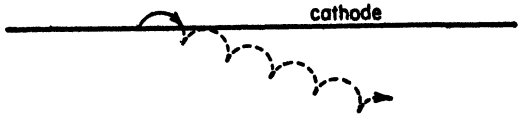


Fig. 10.19. Path of an unfavorable electron, in a region such as (b) or (d) of Fig. 10.17, for a traveling-wave magnetron. Dotted line shows the path if the cathode were not there.

For sustained oscillations to occur, as the favorable electrons start to enter a region such as (b) or (d) in Fig. 10.17, the polarity of the voltage between segments must reverse. Otherwise, they would become energy-removing electrons. Hence, the condition for oscillation is that the left-to-right velocity of the electrons must be such that they travel the distance between anode gaps in one half period of the *RF* voltage, in any number of cycloidal hops desired. That is, if

d = distance between centers of the air gaps
v_0 = left-to-right velocity of the rolling wheel
T = period of the oscillating field

then for oscillations to be sustained,

$$d = v_0\left(\frac{T}{2}\right) = \frac{v_0}{2f}$$

or, solving for the frequency,

$$f = \frac{v_0}{2d}$$

However, it was previously shown that the velocity of the rolling wheel was

$$v_0 = \left(\frac{e}{m}\right)^2 \frac{E}{B}$$

Therefore, the equation for the frequency is

$$f = \left(\frac{1}{2d}\right)\left(\frac{e}{m}\right)^2 \frac{E}{B}$$

Since the frequency depends upon the distance of electron travel, it is definitely a transit time oscillator. Make no mistake, however, there is *no* especial relationship between electron cycloid frequency and tank circuit frequency.

10.11 Electron Bunching (Phase Selection)

Because of the characteristics of this mode of oscillation, a form of electronic bunching occurs similar to that in a Klystron. The bunching tends to occur about the most favorable electron and therefore increases efficiency considerably.

The *most* favorable electron passes through the air gap and interaction field at a time when the field is a maximum. Then, assuming

the period of the cycloidal hop to be small compared to the *RF* period, the vertical components of the *RF* field are equal and opposite on either side of the cycloid. In other words, since the cycloid of the most favorable electron is right in the center of the anode slot, the total vertical electric field producing the cycloid averages out to be the *DC* field.

However, consider an electron which reaches the peak of its hop in region (c) of Fig. 10.17, but to the left of the most favorable electron. The average value of the electric field working on the electron over one of its hops is greater than the steady field and it tends to accelerate the electron, giving it a larger v_0 velocity. Hence, it tends to overtake the most favorable electron. On the other hand, an electron to the right of the center of the air gap tends to be decelerated so that the most favorable electron catches up with it. Thus, conditions are such as to bring relatively unfavorable electrons into favorable phase, greatly increasing efficiency. In effect, the action increases the number of favorable electrons.

This bunching action, and the use of small cycloidal hops, give the traveling-wave magnetron much higher efficiencies and power outputs than can be obtained with cyclotron modes of oscillation. No phase advancement of the bunches of favorable electrons occur so that no trick schemes of removing electrons are required. The operation is simpler because there is no critical relationship between electron period and *RF* period. All of these factors combine to allow power outputs up to 2 and 3 KW of CW power at 50 per cent efficiency in the frequency range around 1800 mcps. Pulse magnetrons (725A) producing an average 80 watts of power (80 KW peak pulse power) at 35 per cent efficiency in the 10,000-mcps (3 cm) band are in general use in service radar transmitters. A 3000-mcps (10-cm) pulse magnetron (720AY) in general use can give peak pulse powers up to 1 megawatt (500 watts average power) at efficiencies up to 60 per cent.

10.12 Space-charge Configuration

Because of this bunching action, the electrons tend to form into clouds, or space charges. In the cylindrical magnetron, the overall electronic mechanism presents a spoke-shaped space charge wheeling around the cathode in synchronism with the anode potential wave. Each spoke of this "electronic wheel" follows the point of maximum retarding field. A typical configuration, which has been confirmed

by calculations, is shown in Fig. (10.20). Within each spoke of the wheel the individual electrons follow the usual epicycloidal paths.

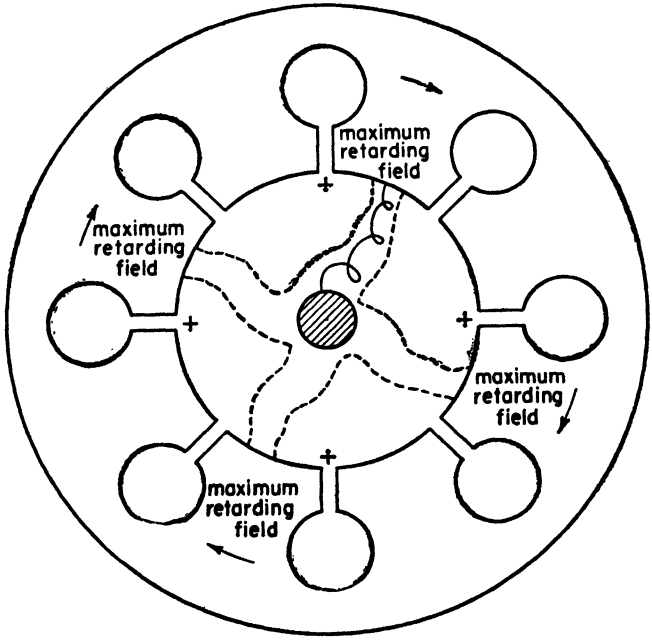


Fig. 10.20. Space charge configuration in a split-anode, resonant-cavity magnetron operating in the traveling-wave mode.

10.13 Modes of Oscillation in the Traveling-wave Magnetron

Because of the high efficiency and large power capacity, magnetrons are almost exclusively operated in the traveling-wave mode. Consequently, split-anode construction is necessarily used. Associated with each pair of anode segments is a tuned circuit which is coupled, largely through mutual inductance, to the other tuned circuits associated with the remaining gaps. This coupling is generally quite close.

It will be recalled that a single resonant circuit has *one* resonant frequency. When two identical tuned circuits are closely coupled, they oscillate at *two* distinct frequencies. This was exemplified by the Klystron. Consequently, in general, the number of resonant frequencies, or modes of oscillation, is a function of the number of tuned circuits. For example, a cylindrical magnetron having six anode segments has six closely coupled tuned circuits associated with

it. Assuming they are all tuned to the same frequency, there are three possible modes of oscillation as shown by the field configurations in Fig. 10.21.

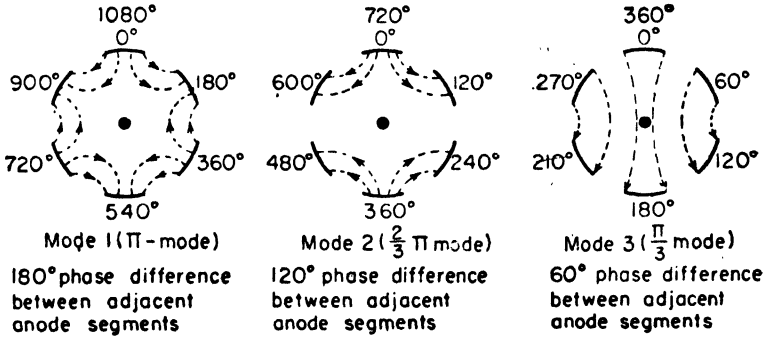


Fig. 10.21. Possible modes of oscillation in a six-anode traveling-wave magnetron.

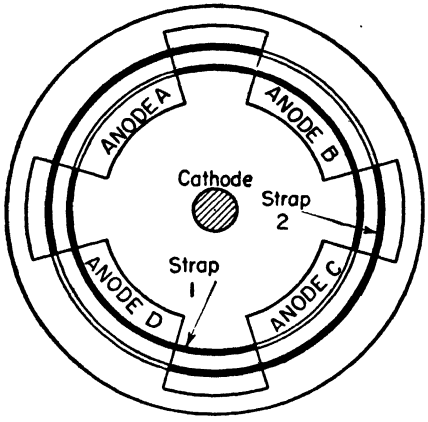
The excitation of mode (1), the π-mode, depends upon adjusting the value of the wheel velocity v_0 so that an electron passes two anode segments in one *RF* cycle. This is the mode discussed in the preceding articles. However, in mode (2), the electron must be made to pass three segments per cycle, while mode (3) requires passing six segments per cycle. This means that modes (2) and (3) require higher velocities of the rolling wheel, v_0. Since v_0 depends upon the ratio E/H, it follows that higher voltages are required to excite the higher modes. Consequently, operation in the π-mode is generally the most desirable and the easiest to produce.

The designation of this magnetron as a traveling-wave magnetron arises from a mathematical analysis which can be performed and which represents the complete field in the magnetron as a rotating field superimposed on a standing wave. Such an analysis enhances computation of electron orbits, but it does not have the physical significance attributable to the discussion as presented here.

10.14 Mode Separation

The frequencies of oscillation in each of the various modes associated with a split-anode magnetron are different. For highest efficiency of transmission, the magnetron should oscillate in only one mode at a time. This condition can be obtained by separating the frequencies of these modes as much as possible so that only one mode at a time is excited.

The original cause of the higher modes was tight coupling between resonant circuits. The tighter the coupling, the greater the separation between resonant frequencies. Hence, one of the general techniques of mode separation is to greatly increase the coupling between resonant circuits. One method of doing this is to connect alternate anode segments (assuming π-mode operation) with metal *straps* (conductive coupling). This holds the polarities of the anode blocks in groups so that, ideally, all anodes in a strapped group follow the same polarity variations. This scheme works effectively until the strap lengths become of the same order of magnitude as the wavelength of the oscillation. At this point, multiple moding again occurs. The method of strapping is indicated in Fig. 10.22. Such a tube is called a *strapped* magnetron.



Strap 1 connects Anodes A and C
Strap 2 connects Anodes B and D

Fig. 10.22. "Strapping" a magnetron for operation in the π-mode.

There is an alternative method of mode separation that depends upon detuning the individual resonant circuits with respect to one another. This method is generally used at the higher frequencies where the strap length would be objectionably large compared to a wavelength. It is produced by using cavities of two different sizes or shapes, as indicated in Fig. 10.23. Because of the physical configuration

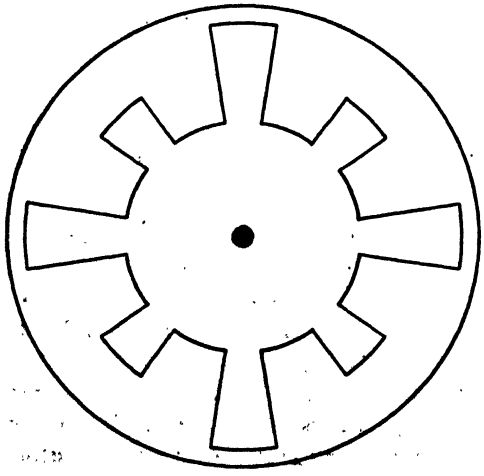


Fig. 10.23. "Rising sun" construction for mode separation.

this is customarily called a *rising sun anode* structure. It has several advantages over strapping, because it is easier to manufacture, does not introduce additional losses, and allows a larger anode structure (better heat dissipation) to be used without penalty in mode separation.

10.15 Typical Cavity Structures

The anode blocks used in magnetrons are usually associated with any one, or any combination, of the following three types of cavities.

(1) Slot type.
(2) Vane type.
(3) Hole-and-slot type.

The last two are less than a quarter wavelength because of the increased capacitive loading at the open ends. The general forms of these cavities are shown in Fig. 10.24. Typical multicavity construction is shown in Fig. 10.23, or 10.20.

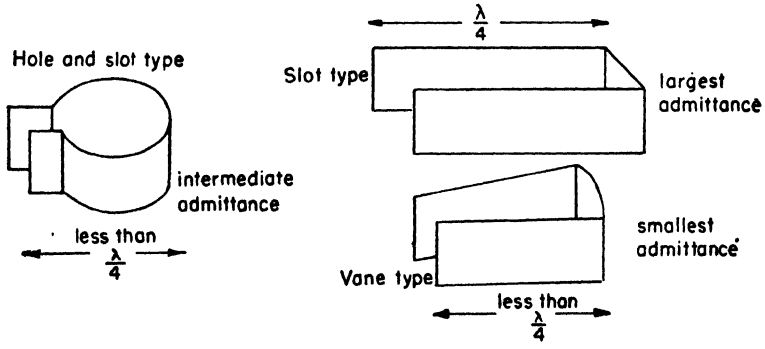


Fig. 10.24. Typical cavity structures used in magnetrons.

10.16 Methods of Tuning

Tuning of a traveling-wave magnetron must be accomplished by a variation in the dimensions of the cavities, or whatever resonant circuits are used, in order to produce a change in susceptance. Tuning magnetrons that use resonant cavities is the most difficult and will be discussed here. It would be possible to change either the capacitance or the inductance, or both, in order to effectuate tuning. As a general rule, the capacitance is varied at the longer wavelengths, while inductive tuning is used at the short wavelengths.

The inductance can be varied by inserting a conducting pin into a hole in each resonator where the lines of magnetic field are con-

centrated. As the pins are pushed in they reduce the volume available for flux, thus reducing the inductance and increasing the frequency. A tuning range of ±7 per cent may be obtained in this manner.

At the longer wavelengths, where strapping is used, another piece of metal shaped like a cookie cutter is placed at the end of the anode block opposite to the straps. It is moved in and out of annular grooves in the block, thus varying the capacitance. Another method, which gives a wider tuning range, is to vary the capacitance of the straps. This is accomplished by adding a metal disk, supported on a tripod, more or less parallel to the strap ring. One leg of the tripod is adjustable so that the separation between strap and tuning ring may be changed at will, thus producing a variation in strap capacitance. This produces a very steep and nonlinear frequency characteristic.

10.17 Mechanical Features

Most pulse magnetrons, and most of the higher-frequency magnetrons, are generally of all metal construction using resonant cavities for tuned circuits. This is the type customarily found in radar transmitters. The CW magnetron, for use at lower frequencies, differs considerably in mechanical detail. It generally uses resonant transmission lines for tuned circuits instead of cavities; it generally has only two anode segments instead of several; it is generally glass enclosed. Furthermore, its mechanical evolution is, to the average reader, more illustrative of engineering technique at work.

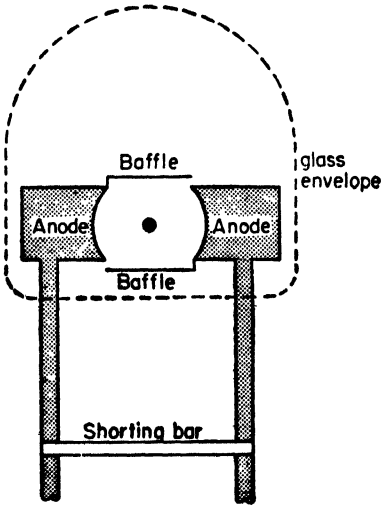


Fig. 10.25. Early form of a CW low-frequency magnetron.

The general form of the CW low-frequency magnetron is shown in Fig. 10.25. The one shown is for use at about 70 to 200 mcps. In order to increase the plate dissipation, water cooling is necessary. This is accomplished by circulating water up through the in-

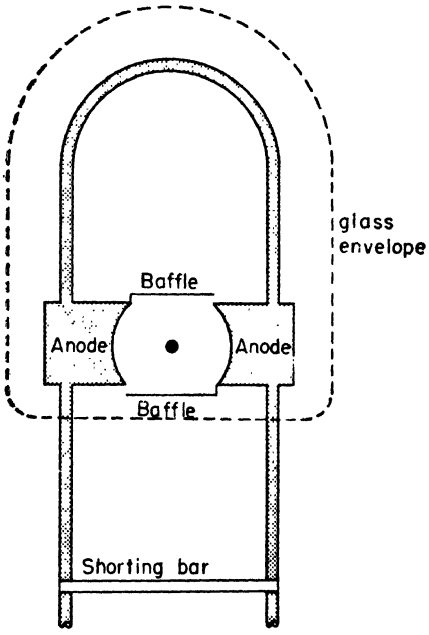


Fig. 10.26. Improved version of the CW magnetron in Fig. 10.25.

side of the tuned lines to the anode block and out again through the same line. Since there is no mechanical connection between anode segments, two separate water systems are required. In the early models of this tube, some electrons escaped through the air gap and proceeded in cycloidal loops, and in a very intense beam, to the glass envelope. Their energy was sufficient to melt the glass. To prevent this, the baffle plates shown were added, with due recognition of the additional capacitance introduced.

The upper frequency limit of this tube is set by the closest point that the shorting bar can be brought up to the tube structure. The extreme limit is generally about $3\lambda/4$. The first major modification intro-

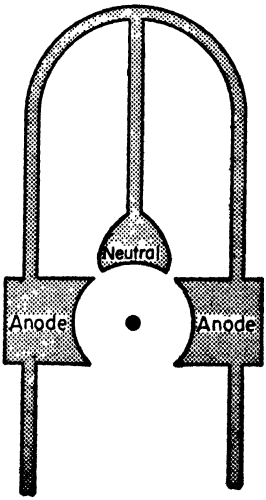


Fig. 10.27. Location of neutral electrode.

duced, aimed at increasing the upper frequency limit, was the addition of the shorted length of transmission line inside the tube

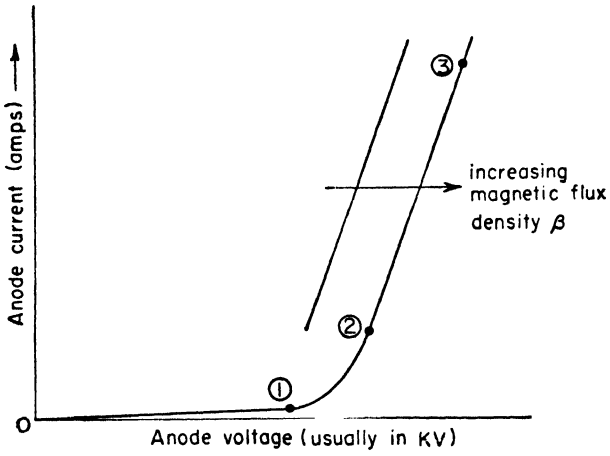


Fig. 10.28. Magnetron characteristics.

envelope, as shown in Fig. 10.26. The anode structure splits the capacitance between the internal and external lines, thus increasing the upper frequency limit. Furthermore, it simplifies the cooling system since one continuous water supply can be used.

A second modification was the addition of a neutral electrode, connected at the neutral point in the short-circuited line as shown in Fig. 10.27. Although there are still only two active anodes, the electron is misled into thinking that there are four.

The original CW magnetron discussed could produce about 150 watts at about 35 per cent efficiency. The addition of the neutral electrode brings these figures up to about 300 watts at 50 per cent efficiency. The most modern tubes of the same general design can deliver 2 or 3 KW at 50 per cent efficiency.

10.18 Magnetron Characteristics

A typical plate current-plate voltage magnetron characteristic is shown in Fig. 10.28. Each of the three regions indicated have different operating characteristics. In the region (0-1), the power output and efficiency are, for all practical purposes, zero. In region (1-2), the power output and efficiency are small. Oscillation tends

to occur in two modes simultaneously. On the straight line part of the curve, region (2-3), the magnetron operates in the π-mode only (if strapped) with a comparatively high efficiency and power output.

REFERENCES

Books

Brainerd, J. G., Koehler, G., Reich, H. J., and Woodruff, I., *Ultra-High-Frequency Techniques*, Van Nostrand, 1942.

M.I.T. Staff, *Principles of Radar*, McGraw-Hill, 1946.

Sarbacher, R. I., and Edson, W. A., *Hyper and Ultrahigh Engineering*, Wiley, 1943.

Periodicals

"Magnetron as a Generator of Centimeter Waves," *Bell System Tech. J.*, Vol. XXV, No. 2, Apr. 1946.

Brillouin, L., "The Theory of the Magnetron," *Phys. Rev.*, **60**, 385 (1941).

Chang, H., and Chaffee, E. L., "Characteristics of the Negative Resistance Magnetron Oscillator," *Proc. IRE* **28**, 519 (1940).

Gill, E. W. B., and Britton, K. G., "The Action of the Split Anode Magnetron," *J. IEE*, **78**, 461 (1936).

Kilgore, G. R., "Magnetostatic Oscillators for Generation of Ultra Short Waves," *Proc. IRE*, **20**, 1741 (1932).

"Magnetron Oscillators for Generation of Frequencies from 300-600 mcps," *Proc. IRE*, **24**, 1140 (1936).

"The Magnetron as a High Frequency Generator," *J. Applied Phys.*, **8**, 666 (1937).

Linder, E. G., "Description of the End Plate Magnetron," *Proc. IRE*, **24**, 633 (1936).

Okabe, K., "Magnetron Oscillations of a New Type," **18**, 1748 (1930).

CHAPTER 11
PROPAGATION OF RADIO WAVES

THE STUDY of communication engineering is facilitated considerably by the manner in which communication systems can be resolved into basic components. In the broadest possible sense, the elements comprising a complete communication system may be grouped into two fundamental classes: terminal devices such as transmitters and receivers, and the medium through which the signal is transmitted in effecting coupling between the two terminations of the system.

In wire communication systems, from the date of their inception, considerable attention was focussed upon the link between the terminal devices, resulting in substantial progress and improvement over the years. Wireless transmission posed a decidedly different problem since the character of the medium of propagation was almost entirely unknown for a long period of time. As a result of Marconi's transatlantic tests about 1900, and of experiments conducted by Westinghouse and a group of radio amateurs who were unimpressed by contemporary scientific doggerel, about 1912, a great deal of interest was directed upon the earth's atmosphere. A considerable amount of study and experiment have occurred in recent years. Progress has been remarkable and the investigation continues.

It is the purpose of this chapter to summarize current knowledge concerning the atmosphere as a medium of radio propagation, and in a generally qualitative, but occasionally quantitative, manner, introduce the student to a new and rapidly growing field of science and engineering.

11.1 Strata Surrounding the Earth

For purposes of analysis in radio propagation problems, the earth's atmosphere, which is the region about the planet containing the air, is usually divided into three principal layers, each of which possesses definite electrical properties, although the layers themselves may not be physically separable in a purely definitive sense. The three layers are the troposphere, stratosphere, and ionosphere.

The *troposphere* is the part of the earth's atmosphere below about six or seven miles where the temperature is a function of the altitude. The troposphere is the weather belt of the earth. The temperature variation, along with changes in the moisture content and density, results in an electrical medium with many nonuniformities that may, under certain conditions, affect electromagnetic transmission considerably.

The *ionosphere*, or Kennelly-Heaviside layer, is that portion of the earth's atmosphere above the lowest level at which the ionization is large compared to that at the ground.

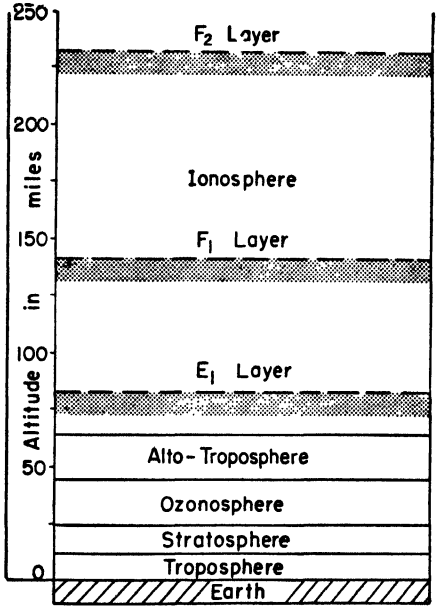


Fig. 11.1. Strata of the earth's atmosphere.

The *stratosphere* is that part of the atmosphere above the troposphere, but below the ionosphere, where the temperature is relatively independent of the altitude. Hence, it is a very nearly homogeneous and isotropic medium of electrical transmission.

The uppermost region of the stratosphere is generally broken up into two other layers called the *ozonosphere* and the *alto-troposphere*. The ozonosphere contains free oxygen, which serves to absorb the actinic rays of the sun. It undergoes a tremendous change in tem-

perature over the 24-hour day. The alto-troposphere also absorbs sunlight and passes through a similarly wide range in temperatures over a 24-hour period. Electrically, the stratosphere, ozonosphere, and alto-troposphere are sufficiently similar to be classified simply as the stratosphere, or isothermal layer. The relative heights and thicknesses of these layers are illustrated by Fig. 11.1.

The presence of the ground (earth, seawater, or whatever it may be) has a marked effect upon propagation. At low frequencies it generally appears as a fairly good conductor, but becomes more dielectric in nature as the frequency is increased.

Since the entire medium of propagation consists of all of these somewhat stratified layers, it is reasonable to expect that the discontinuities in dielectric constant existing at the boundaries between these zones will alter the radiated wave considerably.

11.2 The Character of the Ionosphere

The ionosphere generally begins at a height of about 60 miles above the surface of the earth and is about 200 miles thick. At such high altitudes the pressure is extremely low, being of the same order of magnitude as that in the ordinary commercial vacuum tube. The ionosphere is so termed because it contains large numbers of free ions, principally electrons. The fact that such a layer exists is due to the ionizing action of radiation from the sun on the gas particles composing the air about the earth. At low altitudes the pressure is high, the mean free path of the particles which are ionized by the sun's radiation is so short that recombination occurs in a very short interval of time. At higher altitudes, the pressure goes down and the mean free path undergoes a consequent increase until a point is reached where a liberated electron travels about for a long time before recombining with an ion. Thus, comparatively large numbers of electrons exist at these high altitudes.

The density of the ions and electrons depends upon a multitude of factors such as sunspot activity, altitude, time of day or night and year, and so on. Generally speaking, though, the ionic concentration will depend upon

(1) The nature and degree of radiation from the sun.
(2) The nature of the atmosphere.

The first factor causes the electron density to vary in a very complex manner with time and sunspot activity. The second factor accounts

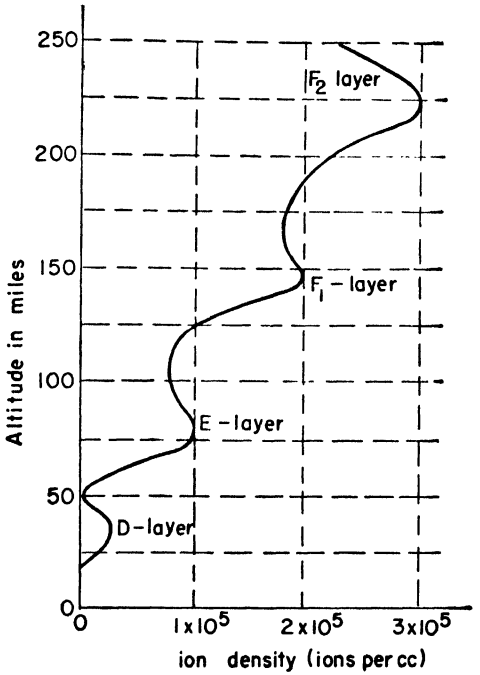


Fig. 11.2. Ionosphere layers during typical daylight hours.

for the dependence of electron density upon altitude, since gas density varies with altitude. Also, the atmosphere is composed of several different gases which do not absorb radiation to the same degree. The relative proportion of these different gases varies with altitude and this is the supposed cause of the stratification of the electrons within the ionosphere.

The tendency toward stratification is shown in Fig. 11.2 which shows the ionic concentration as a function of altitude during a typical daylight period. Figure 11.3 shows typical nighttime conditions.

The D-layer is occasionally present in sufficient strength during day-

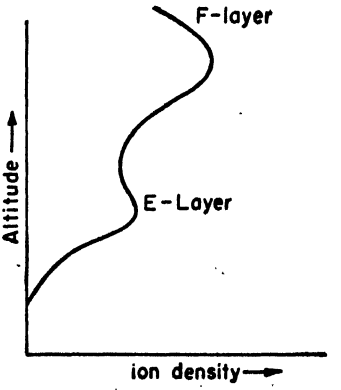


Fig. 11.3. Ionosphere layers during night hours.

light to produce reflections, but it practically disappears at night. The E-layer is consistently strongly ionized, but is subject to strong transient ionic density variations. The F_1-layer is the one principally affected by sunspots and its density may vary by as much as 100 per cent over the 11-year sunspot cycle. At night the D-layer practically disappears and the F_1- and F_2-layers tend to merge into a single stratum producing a curve that characteristically appears as shown in Fig. 11.3. The merging of the F_1- and F_2-layers is largely produced by lowering of the F_2-layer.

The variations in layer height with respect to time are shown by the typical plot in Fig 11.4.

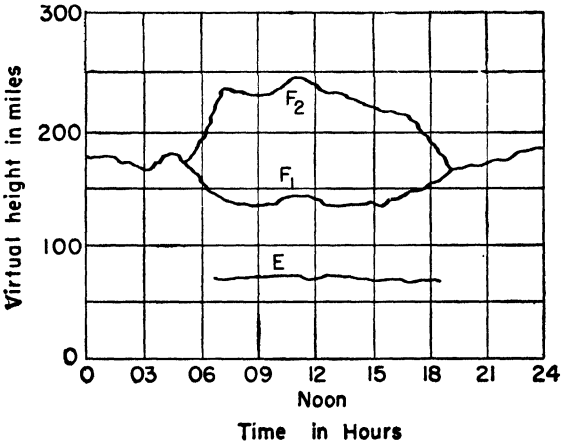


Fig. 11.4. Typical variations in layer height as a function of time.

11.3 Character of the Radiated Wave

In the most general case, the transmitted wave may be analyzed into three major components, the ground wave, a sky wave, and a tropospheric wave. The *ground wave* is that portion of the radio wave which is propagated through space and is affected by the presence of the ground. It is usually subdivided into three components:

(1) Direct wave—that part of the ground wave traveling directly from transmitter to receiver.

(2) Ground-reflected wave—that part of the ground wave reflected by the earth as a result of the dielectric discontinuity existing at the air-to-ground boundary.

(3) Ground-guided wave—that part of the ground wave that is incident upon the earth, and guided and refracted by the air-to-ground interface so that it travels along the surface of the earth.

The *sky, or ionospheric, wave* is the second major component of the total signal radiated from the transmitting antenna. It is that part of the total field which is directed toward, or reflected from the ionosphere. Any part of the signal that is reflected because of a

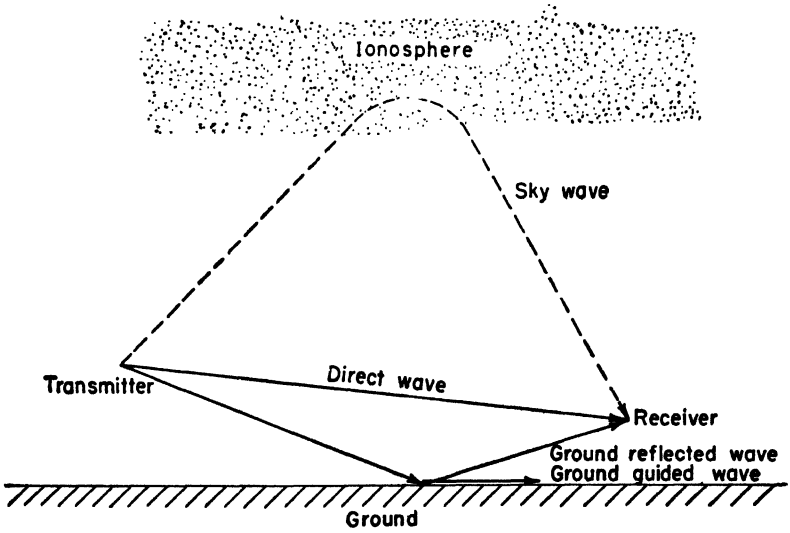


Fig. 11.5. Components of the radiated wave.

dielectric discontinuity in the troposphere is called the *tropospheric wave* and is the last of the major subdivisions of the total signal. Unless definitely specified otherwise in the material that follows, it will be assumed that the tropospheric wave is negligible in comparison to the other components. In other words, it is assumed that the troposphere is homogeneous.

Thus, the components of the radiated wave may be identified as shown by Fig. 11.5. This figure presents a specialized picture, since it shows only those parts of the radiated wave components that appear at the receiver. Actually, there are a large number of such rays, as shown in Fig. 11.6, some of which strike the receiver, and some of which do not.

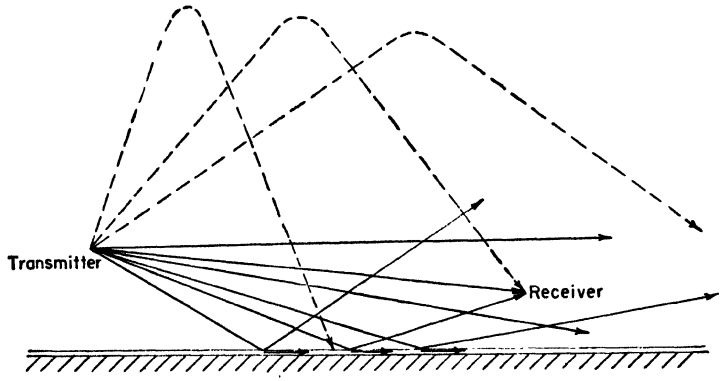


Fig. 11.6. Components of the radiated wave.

11.4 Effect of Frequency on the Components of the Radiated Wave

Certain general propagation phenomena may be ascribed to various frequency bands, so that for the purpose of analysis and discussion, the radio frequency spectrum is usually subdivided as follows:

Band	Wavelength in Meters	Frequency in Kilocycles per Second
Long wave	30,000 - 600	10 - 500
Broadcast	600 - 200	500 - 1500
Short waves—Band A	200 - 100	1500 - 3000
Short waves—Band B	100 - 50	3000 - 6000
Short waves—Band C	50 - 10	6000 - 30,000
Quasi-optical waves	10 - 1	30,000 - 300,000
Microwaves	1 - 0.01	300,000 - 100,000,000

Long waves. Long waves obey the Austin-Cohen formula, derived a great many years ago, which assumes that the wave is propagated through a waveguidelike structure formed by the ionosphere (particularly the D-layer, and sometimes the E-layer), and the conducting earth. There is no separation of the wave into ground and sky wave components; it travels as a single ray. This characteristic renders it eminently well-suited to radio compass applications. Reliable long distance communication may be obtained at these frequencies, but the excessively cumbersome and expensive size of the antenna instal-

lation renders it impractical according to current engineering techniques.

Broadcast waves. During the day, nearly all of the energy that appears at the receiving antenna is due to the ground wave. Hence, practically all broadcast antennas are designed to concentrate the transmitted energy at a low angle of radiation. Nearly all of the high-angle radiation is absorbed by the ionosphere, probably in the D-layer.

At night, the ionization of the D-layer decreases to the point where it is negligible, and the strongly ionized E-layer becomes important by producing strong reflections so that the sky wave becomes a definite consideration.

The intensity of the ground wave diminishes much more rapidly than does the intensity of the sky wave. At distances up to about 50 or 100 miles, the ground wave is stronger than the sky wave. Thereafter, however, the sky wave is comparatively more important. As a result, a zone is reached where the two waves are of about the same intensity. A slight change in atmospheric conditions will produce appreciable changes in the phase relations of the sky wave with respect to the ground wave, resulting in violent fading.

At extreme distances from the antenna the ground wave is negligible, although, under favorable conditions, excellent sky wave transmission may be obtained.

Short waves—Band A. This is a transitional frequency in which the ground wave is so severely attenuated that it is inefficient for broadcast use, and the frequency is too low for effective sky wave communication. Erratic sky waves are produced due to simultaneous reflections from both the E- and F_1-layers, or due to reflections from either one separately. The inconsistency of the reflections and the attenuation of the ground wave render either sky or ground wave transmission inefficient and unreliable, relative to other bands.

Short waves—Band B. This band is widely used by the Armed Forces and the airlines. The rate of attenuation of the ground wave is such that it may be used for reliable transmission over distances up to about 30 miles for soil with high conductivity. Thereafter, however, it is too weak for reception. Except in the region specified, reliance is chiefly placed upon dependable F-layer reflections, although transient E-layer reflections are quite common. These transient reflections, because of sudden and decisive changes in ionic concen-

tration of the E-layer, usually occur so rapidly that the shift from E to F, or from F to E reflections is accomplished with little fading and seldom interrupts service.

Short waves—Band C. The ground wave is rapidly attenuated and within such a short distance that it is not of any practical importance. F-layer reflections occur with a high degree of efficiency so that this band is largely used for long distance communication. It is in this band that the *skip distance* is of such importance. The distance between the end of the importance of the ground wave and the point of reception of the first reflected sky wave is called the *silent zone.* The skip distance is the distance between the point of transmission and the point of reception of the first downcoming sky wave. The area between the first and last downcoming sky waves, at the ground, is the region served by *secondary coverage.* The ground wave supplies the *primary coverage.* These terms are illustrated by Fig. 11.7.

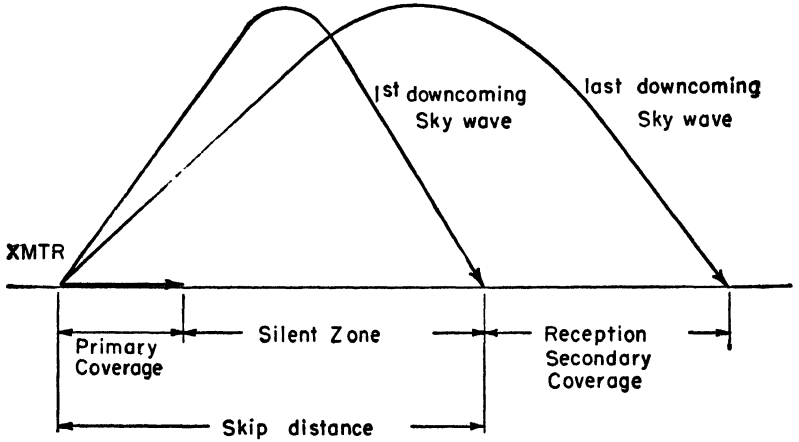


Fig. 11.7. Skip distance phenomenon.

The skip distance phenomenon also occurs in Band B, but its effect is largely obscured by the ground wave since the region of primary coverage extends practically to the point of reception of the first downcoming sky wave.

Increasing the frequency increases the skip distance until, at a critical frequency, around 10 mc, and under average conditions, the sky wave penetrates the ionosphere, is not returned to the earth, and the

silent zone embraces the earth. This makes sky wave transmission impossible. Slightly below this critical frequency, extremely long-distance communication may be obtained. Slightly above it, no effective transmission occurs. The critical frequency may vary over a considerable range owing to changes in the ionosphere so that a very long-range transmission is usually produced at frequencies close to this critical frequency, but far enough removed from it to make transmission reliable.

Quasi-optical and microwaves. The frequency is too high for use of the sky wave, and the ground wave is promptly attenuated. Thus, transmission occurs largely over optical distances by means of the direct wave. In this frequency band, variations in the index of refraction of the troposphere produce a certain amount of refraction of the direct wave, causing it to be propagated over longer distances than would be expected from ordinary line-of-sight calculations.

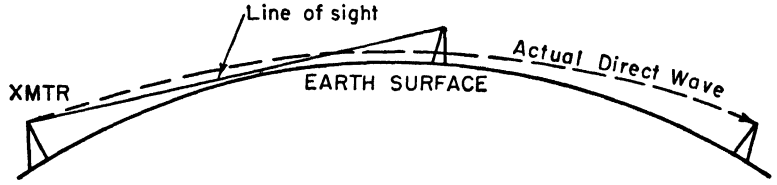


Fig. 11.8. Effect of refraction in the troposphere.

Tropospheric discontinuities caused by extreme weather conditions produce appreciable reflections at certain frequencies. This phenomenon was exploited during the past war by meteorologists who effectively used 3-cm radar sets as "storm hunters," depending upon sufficiently strong reflections from clouds to locate storm centers.

11.5 Effect of Ground on Transmission

In the frequency band in which amplitude-modulated broadcasting is used, it was pointed out in the preceding article that sky wave transmission during the daylight hours is not feasible. Consequently, the service area for a broadcasting station is limited to the distance that it can transmit the ground wave with sufficient signal strength to drive a $14.95 superheterodyne receiver. The usual area of adequate signal strength is generally assumed to be within the 500-μvolt contour.

Good engineering will produce maximum signal strength in the immediate vicinity of the transmitter, but thereafter, nature runs its course. The ground introduces a power loss to the radiated wave that is a function of the soil conductivity and the operating frequency of the station. For a given location, the conductivity is relatively fixed; at least, there is not much that the station engineers can do

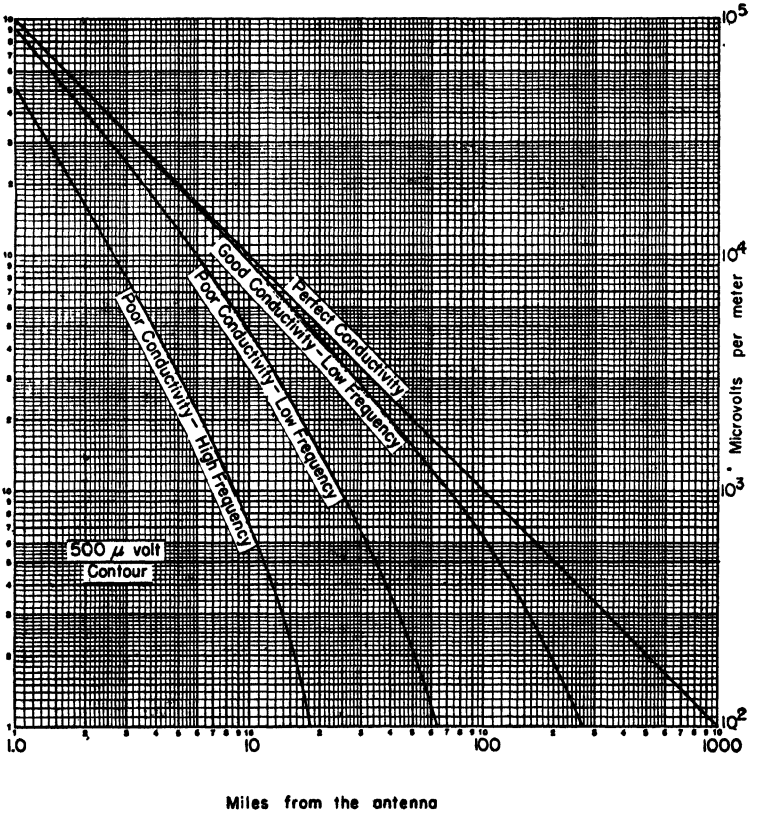


Fig. 11.9. Effect of ground on broadcast station service area.

about it. But, it is interesting to ascertain the effects of frequency on the service area, assuming two extremes of soil conductivity. The results appear generally as shown by Fig. 11.9. Assuming perfect conductivity for the earth, the field strength as a function of distance appears as shown by the straight line in the figure. This relationship

would hold for any frequency, but it would be affected by the power of the station. At a low frequency (around 550 kcps) and with a good conducting earth, the 500-μvolt contour extends to over 100 miles. A marked drop in soil conductivity contracts the service radius to about 35 miles. But, for this lower conductivity the service radius drops to about 10 miles if the frequency is increased to the high end of the broadcast band (about 1500 kcps). It should be obvious then that ground wave transmission, at frequencies above the broadcast band, is economically impractical.

11.6 The Dielectric Constant of the Ionosphere

The presence of free charges in considerable quantity in the ionosphere give it distinct and unique properties as a medium of electromagnetic wave propagation. The effect of the presence of these charges is most conveniently accounted for by calculating a new dielectric constant for the ionosphere that can be expressed in terms of the dielectric constant for free space and properties and densities of the free charges. For the present, the effect of the earth's magnetic field will be omitted from the calculation.

Assume a dielectric medium of dielectric constant ϵ, which contains a gas of free charges, each of charge q. If an electric field $\mathbf{E}$ is applied across a volume of this medium, the force on the charges due to this field may be written as

$$\mathbf{f} = q\mathbf{E} = m\mathbf{a} \tag{11.1}$$

where m is the mass of the charge. Replace the acceleration $\mathbf{a}$ by $d\mathbf{v}/dt$, obtaining

$$q\mathbf{E} = m\frac{d\mathbf{v}}{dt} \tag{11.2}$$

Assume that the electric field intensity is a harmonic function of time according to the relationship below:

$$\mathbf{E} = \mathbf{E}_m \sin \omega t \tag{11.3}$$

Substituting back into Eq. (11.2) yields

$$q\mathbf{E}_m \sin \omega t = m\frac{d\mathbf{v}}{dt} \tag{11.4}$$

or, solving for the differential velocity,

$$d\mathbf{v} = \frac{q}{m}\mathbf{E}_m \sin \omega t \, dt \tag{11.5}$$

Integrate both sides of the equation, obtaining

$$\mathbf{v} = -\frac{q\mathbf{E}_m}{\omega m}\cos \omega t + K \tag{11.6}$$

The constant of integration, K, can be neglected because interest is centered on the *variation* in velocity, not in its absolute value.

The general expression for the current in any medium contains two separate terms,

$\mathbf{i}$ = conduction current + displacement current

$\mathbf{i} = \mathbf{i}_1 + \mathbf{i}_2$

The displacement current, $\mathbf{i}_2$, is

$$\mathbf{i}_2 = \frac{\partial \mathbf{D}}{\partial t} = \epsilon \frac{\partial \mathbf{E}}{\partial t} = \omega \epsilon \mathbf{E}_m \cos \omega t \tag{11.7}$$

The conduction current, $\mathbf{i}_1$, is

$$\mathbf{i}_1 = Q\mathbf{v} = Nq\mathbf{v} \tag{11.8}$$

where N is the number of charges per unit volume, q is the charge, and $\mathbf{v}$ is the charge velocity. From Eq. 11.6, the velocity is given as

$$\mathbf{v} = -\frac{q\mathbf{E}_m}{m\omega}\cos \omega t$$

which, when substituted in the equation for the conduction current, gives

$$\mathbf{i}_1 = -\frac{Nq^2}{\omega m}\mathbf{E}_m \cos \omega t \tag{11.9}$$

Consequently, the expression for the total current is

$$\mathbf{i} = \left(\omega \epsilon - \frac{Nq^2}{\omega m}\right)\mathbf{E}_m \cos \omega t \tag{11.10}$$

Rearranging terms,

$$\mathbf{i} = \left(\epsilon - \frac{Nq^2}{\omega^2 m}\right)\omega \mathbf{E}_m \cos \omega t \tag{11.11}$$

However, the last factor, $\omega \mathbf{E}_m \cos \omega t$, is actually the derivative of the electric field intensity. That is,

$$\omega \mathbf{E}_m \cos \omega t = \frac{\partial \mathbf{E}}{\partial t} \tag{11.12}$$

Hence,

$$\mathbf{i} = \left(\epsilon - \frac{Nq^2}{\omega^2 m}\right)\frac{\partial \mathbf{E}}{\partial t} \tag{11.13}$$

The ordinary expression for displacement current is

$$\mathbf{i} = \epsilon' \frac{\partial \mathbf{E}}{\partial t} \tag{11.14}$$

Consequently, from a term-by-term comparison of Eqs. (11.13) and (11.14), it can be deduced that the medium under consideration can be thought of as a dielectric medium with a dielectric constant ϵ', where

$$\epsilon' = \left(\epsilon - \frac{Nq^2}{\omega^2 m}\right) = \epsilon\left(1 - \frac{Nq^2}{m\omega^2\epsilon}\right) \tag{11.15}$$

The medium on which the analysis was made is analogous to the ionosphere. Hence, the ionosphere acts like a medium with a dielectric constant ϵ' such that

$$\epsilon' = \epsilon\left(1 - \frac{ne^2}{m\omega^2\epsilon}\right) \tag{11.16}$$

where ϵ' is the dielectric constant of the ionosphere, ϵ is the dielectric constant of free space, n is the number of free electrons in space per unit volume, m is the mass of the electron, e is the charge on the electron, and ω or $2\pi f$ is the angular velocity of the electric field.

Because of the dependence of ϵ' on the frequency f, there is some value of f (or ω), at which the dielectric constant of the ionosphere is zero. This is usually called the *critical frequency* and denoted by f_c. Letting $f = f_c$ in Eq. (11.16) yields

$$f_c^2 = \frac{ne^2}{4\pi^2\epsilon m} \tag{11.17}$$

Substituting Eq. (11.17) back into Eq. (11.16) yields

$$\epsilon' = \epsilon\left(1 - \frac{f_c^2}{f^2}\right) \tag{11.18}$$

Apparently then, as a function of frequency, the dielectric constant of the ionosphere can vary over a range from the dielectric constant of free space, through zero, to minus infinity.

11.7 Index of Refraction of the Ionosphere

The equation for the dielectric constant of the ionosphere, Eq. (11.18), may be used to obtain the index of refraction of the ionosphere. To a first approximation, the problem at the boundary

between the stratosphere and ionosphere may be represented as shown in Fig. 11.10 in which it has been assumed that a sharp boundary exists between the two strata. φ_I is the angle of incidence of the radiated sky wave on the boundary; φ_T is the angle of the transmitted wave, that is, the refracted wave; φ_R is the angle of reflection, which is the same as the angle of incidence.

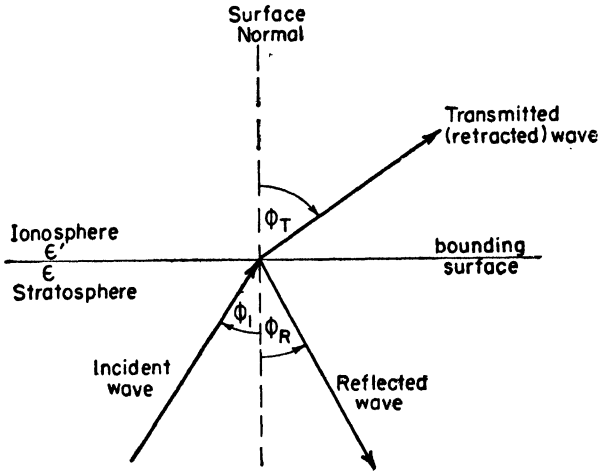


Fig. 11.10. Hypothetical reflection at boundary between ionosphere and stratosphere.

The phase velocity v of a wave in a homogeneous isotropic medium is given by the following expression:

$$v = \sqrt{\frac{1}{\mu\epsilon}}$$

where μ is the permeability and ϵ, the dielectric constant. The index of refraction at the boundary between two different, but homogeneous, mediums, is given by

$$\eta = \frac{v_s}{v_i}$$

where η is the index of refraction, v_s is the phase velocity in one medium (stratosphere), and v_i is the phase velocity in the other medium (ionosphere). In the case under consideration the permeability is the same in both mediums, so that substitution of the

expressions for v_s and v_t into the equation for the index of refraction yields

$$\eta = \sqrt{\frac{\epsilon'}{\epsilon}} = \sqrt{1 - \left(\frac{f_c^2}{f^2}\right)} \tag{11.19}$$

Thus, it is apparent that the index of refraction, like the dielectric constant, is a function of frequency. When this equation was first derived, several years ago, considerable controversy resulted, since the index of refraction η becomes imaginary when $(f_c/f) > 1$. The meaning of an imaginary index of refraction was the subject of widespread discussion. Subsequently, a more exact analysis by Pederson showed that, actually, η was always real and given by the relation

$$\eta = \left[\frac{\epsilon'}{2\epsilon} + \sqrt{\left(\frac{\epsilon'}{2\epsilon}\right)^2}\right]^{1/2} \tag{11.20}$$

where

$$\epsilon' = \epsilon\left(1 - \frac{f_c^2}{f^2}\right)$$

Consequently, when the dielectric constant of the ionosphere is positive, the index of refraction is

$$\eta = \sqrt{\frac{\epsilon'}{\epsilon}}$$

as given by Eq. (11.19). But, when ϵ' is negative, the index of refraction is zero. Hence, Eq. (11.19) is true for positive values of ϵ' only.

11.8 The Maximum Usable Frequency (MUF)

For maximum efficiency of transmission, that is, for maximum signal at the receiving antenna, it would be highly desirable to produce total reflection from the ionosphere. For total reflection, the angle of the refracted wave must be 90° as measured from the normal to the boundary. In general, from optics it is known that the index of refraction can be specified in terms of the angles of the incident and refracted waves, as given in the following equation.

$$\eta = \frac{\sin \varphi_I}{\sin \varphi_T} \tag{11.21}$$

Therefore, if the conditon for total reflection is assumed,

$$\varphi_T = 90°$$

and hence,

$$\sin \varphi_T = 1$$

Therefore, the equation for the index of refraction is

$$\eta = \sin \varphi \tag{11.22}$$

But, from Eq. (11.19)

$$\eta = \sqrt{1 - \left(\frac{f_c}{f}\right)^2} \tag{11.23}$$

Equating (11.22) and (11.23) gives

$$\sin \varphi_I = \sqrt{1 - \left(\frac{f_c}{f}\right)^2} \tag{11.23a}$$

Square both sides, collect terms, and solve for the frequency f This process yields

$$f^2 = \frac{f_c^2}{1 - \sin^2\varphi_I} = \frac{f_c^2}{\cos^2\varphi_I} = f_c^2 \sec^2\varphi_I \tag{11.24}$$

or

$$f = f_c \sec\varphi_I \tag{11.25}$$

Hence, for all values of the angle of incidence of the radio wave greater than that given by Eq. (11.23a), total reflection occurs, and Eq. (11.25) gives the maximum usable frequency (MUF) for which total reflection can be produced. Thus, for most effective use of the ionosphere in transmission, the radio frequency must not exceed the MUF.

11.9 Virtual Height of the Ionosphere

The discussion thus far has proceeded upon the inference that the boundary separating the ionosphere from the rest of the atmosphere

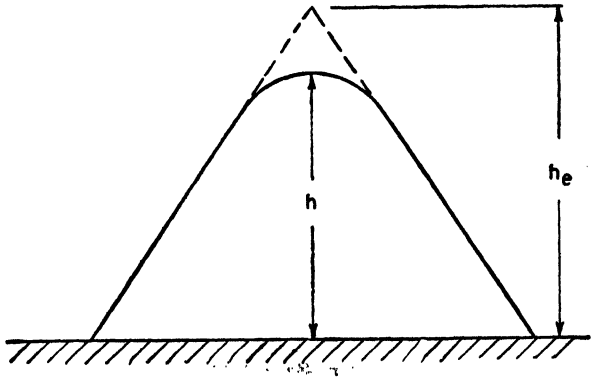


Fig. 11.11. Virtual height of the ionosphere.

is a sharp discontinuity in dielectric constant, so that reflection occurs at the boundary surface. Actually, the line of demarcation is not sharp, but changes rather gradually as is indicated by Fig. 11.2, which shows the electron density as a function of altitude. In such a case, true reflection does not occur; the incident wave is gradually refracted away from the normal to the boundary at a rate depending upon the gradational increase of electron density. If this rate of increase is sufficient, the total refraction will eventually reverse the ray direction and send it downward, as indicated by Fig. 11.11.

The refraction results from an increase in phase velocity, not energy velocity, owing to the change in dielectric constant with increasing electron density. The problem in geometry is considerably simplified if it is assumed that the ionosphere actually does present a sharp boundary at the height h_e shown in Fig. 11.11. The approximation yields *equivalent*, or *virtual*, *heights* h_e consistent with experimental results.

The virtual height may be measured by a method similar to radar and originally conceived by Breit and Tuve about 1926. A pulse is transmitted vertically up, and the time of travel of the returned echo is measured. When converted to distance, assuming the *RF* pulse velocity is the same as the velocity of light, a graph of the form shown in Fig. 11.12 is obtained. In such a case, the angle of incidence is 0°, so that, from Eq. 11.25

$$f = f_c \sec \varphi_I = f_c \qquad \text{Since sec } 0° = 1$$

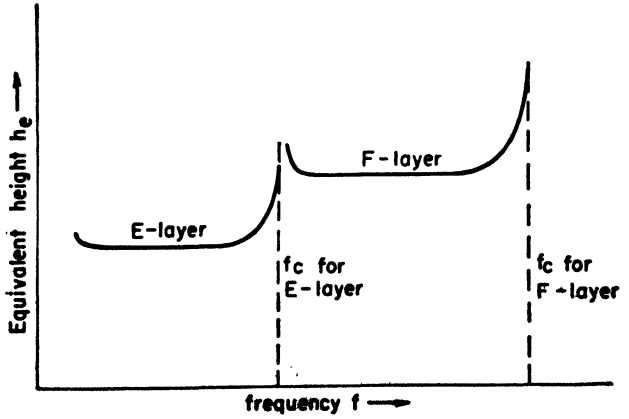


Fig. 11.12. Experimentally determined layer curves.

Hence, the critical frequency f_c can be defined as the maximum frequency which will produce total reflection from a vertically incident wave. The virtual height h_e is the height at which reflection would occur in a purely geometrical manner from a sharp surface discontinuity as shown by the dotted line in Fig. 11.11.

11.10 Practical Application of Ionospheric Reflection

The application of this material to a practical problem in radio transmission presents a rather neat geometrical analysis, which is accomplished by the engineer's stand-by, cut and try. The problem posed is essentially this:

> A transmitter operating at a particular location is required to lay down a signal at some designated receiving point, so far distant that sky wave transmission is necessitated. The values of operating frequency and angle of incidence of the transmitted wave with the ionosphere that must be used to accomplish the transmission, are the factors to be determined.

It was previously shown that the maximum usable frequency is given by

$$f = f_c \sec\varphi_I$$

Assuming a plane earth and disregarding the polarizing effects of the earth's magnetic field, the problem in geometry is as shown in Fig. 11.13. Assume an operating frequency f_x, and then assume a series

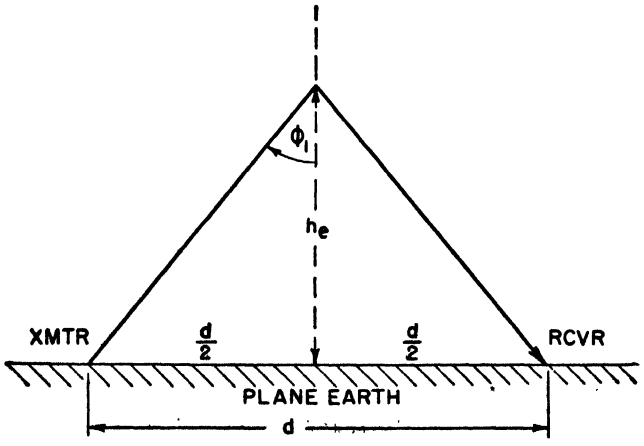


Fig. 11.13. Geometry of sky-wave transmission.

of values for the virtual height which, from experience, are heights at which reflection may reasonably be expected to occur. Since the distance of transmission d and virtual height h_e are then known, values of the angle of incidence φ_I may be obtained by geometry for each value of equivalent height assumed. Substituting the values of the angle of incidence so calculated into the equation for the frequency, given above, yields

$$f_x = f_c \sec \varphi_I$$

This yields a series of values for critical frequency f_c which correspond to the various values of virtual height assumed. Repeat this process for several other assumed operating frequencies, f_y and f_z. If the values of critical frequency obtained by this procedure are plotted against the assumed values for the virtual height, the results appear as shown by the dotted curves in Fig. 11.14. The solid

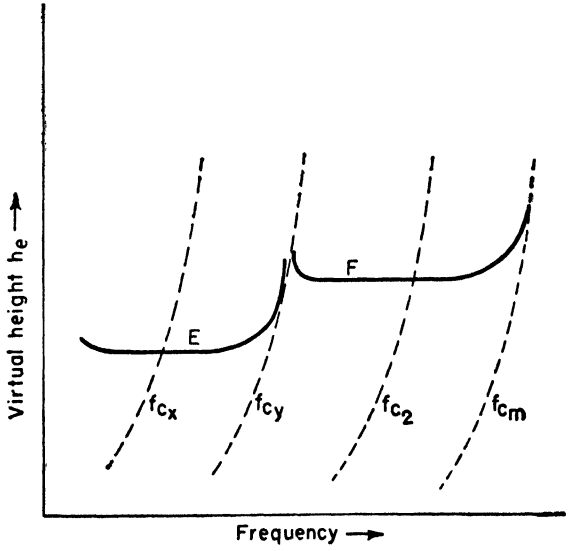


Fig. 11.14. Method of solution of sky-wave transmission problem.

curves are actual experimentally determined plots of the equivalent height as a function of frequency.

The curve for critical frequency corresponding to the assumed operating frequency f_x is f_{c_x}, and, as drawn in Fig. 11.14, shows that

it is a possible frequency for E-layer transmission, just as the intersection of f_{c_z} with the F-layer shows that f_z is a possible operating frequency for F-layer transmission. f_{c_y} is obtained from the assumed frequency f_y, and since it does not intersect either layer, f_y is an operating frequency for which no sky wave transmission would be feasible. The selection of the operating frequency f_m produces a curve of critical frequency f_{c_m} that is just tangent to the F-layer and thus represents the F-layer MUF.

In general, it is highly desirable to obtain the value of the MUF for each layer since the selection of the operating frequency is based upon the value of the MUF. If the operating frequency is much less than the MUF, the losses in the ionosphere become large and efficiency is reduced. On the other hand, if the operating frequency is very close to the MUF, slight changes in the ionosphere (which always occur) will make transmission very erratic. Consequently, it is common practice to strike a compromise and operate at 85 per cent of the MUF. The resulting frequency is usually called the *optimum working frequency* (OWF).

Now, return to the problem. Since f_m is the MUF for the F-layer, the station operating frequency should be picked as 85 per cent of f_m so as to use the OWF for that layer. Using the OWF as the operating frequency, follow the same procedure as before, obtaining a plot of critical frequency corresponding to the OWF. The intersection of this curve with the experimental-layer curve gives the value of virtual height h_e from which the required angle of incidence can be calculated.

It is possible that the plotted curve for critical frequency may intersect the experimental-layer curves at more than one point. This produces dual reflections, which arrive at the receiver with different amplitude and phase conditions, resulting in frequent and appreciable fading. Consequently, it would not be a desirable operating frequency.

At the present time the data from which such calculations as these are made are compiled and prepared into nomographs by the Bureau of Standards. This enhances computation, but it obscures the fundamental nature of the analysis. When the student is familiar with the underlying mechanics of the problem, he is equipped to undertake the use of the prepared nomographs.

11.11 Effect of the Earth's Magnetic Field

The problem of electromagnetic propagation in the ionosphere has been somewhat oversimplified by neglecting the effect of the earth's magnetic field. The presence of this field complicates transmission and causes deviations from the simple theory enunciated so far.

Consider a plane electromagnetic wave being propagated through the ionosphere. Assume the coordinate system in such a way that the direction of propagation of the wave is along the Z axis. Since propagation is in the Z direction, and since the wave is a plane wave, then the electric and magnetic fields are entirely within the X-Y plane and neither field has a Z component.

Assume that there are n particles of charge e and mass m per unit volume. To account for the earth's magnetic field, assume that a static magnetic field is present that may be resolved into two components, one transverse to the direction of propagation and one parallel (longitudinal) to the Z axis. Only the longitudinal component H_e will be considered since it is the only one that produces a primary effect on the traveling wave.

The force on any one of the charges may be written as

$$\mathbf{f} = m\mathbf{a} = m\frac{d^2\mathbf{s}}{dt^2} \tag{11.26}$$

In terms of the field quantities, both the magnetic and electric fields exert forces. That is

$$\mathbf{f} = e\mathbf{E} + e(\mathbf{B} \times \mathbf{v}) \tag{11.27}$$

where $\mathbf{E}$ is the electric field intensity and $\mathbf{H}$ is the magnetic field intensity $= \mathbf{B}/\mu_0$. Assuming the section of the ionosphere under consideration to be homogeneous and isotropic, then

$$\mathbf{f} = e\mathbf{E} + \mu_0 e(\mathbf{H} \times \mathbf{v}) \tag{11.28}$$

Equating the two expressions for the force on the charge,

$$\frac{d^2\mathbf{s}}{dt^2} = \frac{e}{m}[\mathbf{E} + \mu_0(\mathbf{H} \times \mathbf{v})] \tag{11.29}$$

Carry out the indicated vector cross product and equate like components of the vectors on either side of the equal sign. This process yields the following equations:

$$\frac{d^2x}{dt^2} = \frac{e}{m}[E_x + \mu_0(v_yH_0 - v_zH_y)] \tag{11.30}$$

$$\frac{d^2y}{dt^2} = \frac{e}{m}[Ey + \mu_0(v_z H_x - H_0 v_x)] \tag{11.31}$$

$$\frac{d^2z}{dt^2} = \frac{e}{m}[\mu_0(v_x H_y - v_y H_x)] \tag{11.32}$$

Note that the longitudinal component of the earth's magnetic field H_0 is the Z component of the *total* magnetic field. The electric field has no Z component at all, of course.

The force exerted by the electric field of the incident wave is, for the X component,

$$f_e = eE_x$$

and the force for the magnetic field is

$$f_m = \mu_0 v_z H_y$$

The ratio of these two forces is

$$\frac{f_e}{f_m} = \left(\frac{e}{\mu_0 v_z}\right)\frac{E_x}{H_y} = \frac{e}{v_z}\left(\frac{1}{\mu_0}\right)\sqrt{\frac{\mu_0}{\epsilon_0}} = e\left(\frac{c}{v_z}\right)$$

where $E_x/H_y = \sqrt{\mu_0/\epsilon_0}$ for a plane wave in free space (see Art. 6.14), c is the velocity of light $(= \sqrt{1/\mu_0\epsilon_0})$, and v_z, the charge velocity. Now, the velocity of light is very much greater than the charge velocity, so the ratio is practically infinite. Hence, the force due to the magnetic field of the traveling wave is negligible compared to the force exerted by the electric field. The equations of motion consequently reduce to

$$\frac{d^2x}{dt^2} = \frac{e}{m}(E_x + \mu_0 H_0 v_y) \tag{11.33}$$

$$\frac{d^2y}{dt^2} = \frac{e}{m}(E_y - \mu_0 H_0 v_x) \tag{11.34}$$

$$\frac{d^2z}{dt^2} = 0$$

Therefore, the charge motion is purely transverse to the direction of wave propagation, that is, motion is entirely within the X-Y plane. A considerable simplification can then be made by using the complex notation of ordinary circuit analysis. In other words, let

$$E = E_x + jE_y \tag{11.35}$$

$$H = H_x + jH_y \tag{11.36}$$

$$s = x + jy \tag{11.37}$$

Therefore Eqs. (11.33) and (11.34) can be written as

$$\frac{d^2s}{dt^2} = \frac{e}{m}\left(E - \mu_0 H_0 \frac{ds}{dt}\right) \tag{11.38}$$

Or, in the customary form for a differential equation,

$$\frac{d^2s}{dt^2} + \left(\frac{e\mu_0 H_0}{m}\right)\frac{ds}{dt} - \left(\frac{eE}{m}\right) = 0 \tag{11.39}$$

The expressions for E and H may be obtained in a general form from the wave equations. In Art. 6.14 it was shown that the electric and magnetic fields in a TEM (plane) wave have the following general character, assuming harmonic time dependence:

$$E = Ae^{\pm j(\gamma z - \omega t)} \qquad H = Be^{\pm j(\gamma z - \omega t)} \tag{11.40}$$

Assume a similar form for the charge displacement, that is, let

$$s = Ce^{\pm j(\gamma z - \omega t)} \tag{11.41}$$

It is a valid assumption because the form of the differential equation for the displacement is substantially the same as the wave equation for a plane wave. These three assumed solutions must satisfy Eq. (11.39) and Maxwell's equations, which are given below.

$$\nabla \times \mathbf{E} = \text{Curl } \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t}$$

$$\nabla \times \mathbf{H} = \text{Curl } \mathbf{H} = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J}$$

For the plane wave under consideration, and in the complex notation specified, these equations reduce to

$$\begin{aligned} \frac{\partial E}{\partial z} - j\mu_0 \frac{\partial H}{\partial t} &= 0 \\ \frac{\partial H}{\partial z} + j\epsilon_0 \frac{\partial E}{\partial t} &= jne \frac{ds}{dt} \end{aligned} \tag{11.42}$$

To verify the validity of the assumed solutions, substitute Eqs. (11.40) and (11.41) into Eqs. (11.39) and (11.42). This process yields

$$\begin{aligned} \left(-\frac{e}{m}\right)A + \left(-\omega^2 \pm \frac{e\mu_0 H_0 \omega}{m}\right)C &= 0 \\ (n)A + (j\mu_0\omega)B &= 0 \\ (e_0\omega)A + (jn)B + (ne\omega)C &= 0 \end{aligned}$$

A trivial solution would result if the constants A, B, and C were zero. So, if the assumed solutions are to be valid, then the value of the third-order determinant formed by the coefficients of the constants must be zero. Setting up this determinant and evaluating it leads to the following equation:

$$\frac{\gamma^2}{\mu_0\epsilon_0\omega^2} = 1 - \frac{ne^2/m\epsilon_0}{\omega^2 \mp e\mu_0 H_0\omega/m} \tag{11.43}$$

But c is the velocity of light $\sqrt{1/\mu_0\epsilon_0}$; v_1, the phase velocity γ/ω; and η, the index of refraction c/v. Hence, substituting for these quantities yields

$$\eta^2 = 1 - \frac{ne^2/m\epsilon_0}{\omega^2 \mp (e\mu_0 H_0\omega/m)} \tag{11.44}$$

This produces two possible values for the index of refraction, η_+ and η_-, depending upon the choice of sign in the denominator. Hence, the ionosphere has two modes of propagation with distinctly different phase velocities corresponding to the dissimilar indices of refraction.

The effect of the earth's magnetic field on a radio wave is rather remarkable. A linearly polarized wave entering the ionosphere is split up into two components, a right-handed and a left-handed circularly polarized wave. The first is called the *extraordinary* wave and is propagated with a phase velocity v_+, while the latter is called the *ordinary* wave and has a velocity v_-.

Reflection of the sky wave at the hypothetical boundary of the ionosphere is determined by the index of refraction. As a result, the wave returning to earth may have a strong circularly polarized component present that contributes to fading.

11.12 Standard Atmospheric Refraction

In the FM and television parts of the frequency spectrum, transmission is accomplished largely over optical distances by the direct and ground-reflected waves. Transmission occurs almost exclusively within the troposphere. Up to this point the troposphere has been assumed to be homogeneous, so that the dielectric constant, and consequently the index of refraction, is constant. Actually, this is not quite true. As the altitude increases, the temperature, pressure, and water vapor content all generally decrease. This causes a gradual reduction in the index of refraction with increasing altitude. Consequently, the phase velocity increases somewhat at the higher

altitudes, causing a very gentle refraction of the direct wave. To a small extent, the wave follows the curvature of the earth and causes transmission to occur over larger distances than would be predicted by ordinary line-of-sight calculations.

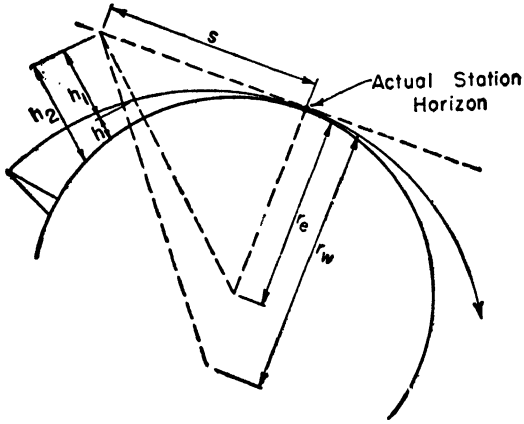


Fig. 11.15. Curved wave over actual earth.

Calculations of station coverage are usually, and more easily, made by assuming straight-line transmission. The effect of atmospheric refraction is then accounted for by assuming the earth to have a larger radius than it actually does. The specification of this effec-

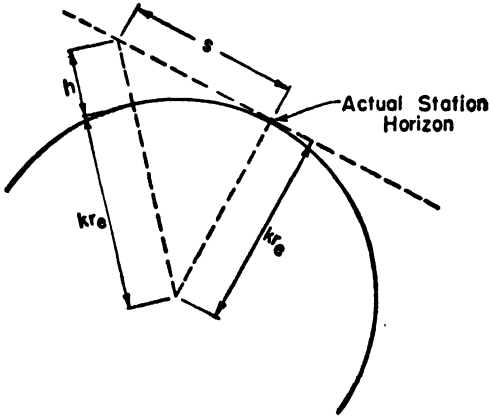


Fig. 11.16. Assumed straight ray over fictitious earth.

tive radius depends upon the character of the troposphere. A standard has been established that is the average atmospheric conditions for the middle latitudes of North America. This is referred to as a *Standard Atmosphere* and the refraction it produces is called *Standard Atmospheric Refraction.*

Assuming a standard atmosphere, the fictitious radius of the earth is selected in such a way that the straight line drawn from the transmitter is tangent to this hypothetical earth at the same distance from the transmitter as the actual refracted signal is tangent to the actual earth. Figure 11.8 shows the extension of the station service area to a new horizon as a result of atmospheric refraction. Figures 11.15 and 11.16 show the relationship between the actual curved ray over the true earth and the assumed straight line over the fictitious earth.

As a matter of terminology, let r_e be the actual radius of the earth; r_w, the radius of the curved path of the actual wave; h, the height of the wave above the earth; s the distance of travel of the wave to the point of tangency with the earth; and k, the constant specifying the proportion between r_e and r_w. From the right triangle in Fig. 11.16,

$$(s^2) + (kr_e)^2 = (h + kr_e)^2$$

or

$$s^2 + (kr_e)^2 = h^2 + 2hkr_e + (kr_e)^2$$

By inspection of the figure it is apparent that

$$h \ll h(kr_e) \qquad \text{since} \qquad kr_e \gg 1$$

Hence, the equation reduces to

$$s^2 = 2hkr_e$$

Solving for h yields

$$h = \frac{s^2}{2kr_e}$$

Following the same procedure for the right triangles of Fig. 11.15,

$$h_1 = \frac{s^2}{2r_e} \qquad \text{and} \qquad h_2 = \frac{s^2}{2r_w}$$

But, from Fig. 11.15, by definition,

$$h = h_1 - h_2$$

Therefore,

$$\frac{1}{kr_e} = \frac{1}{r_e} - \frac{1}{r_w}$$

Solving for the constant k yields

$$k = \frac{r_w}{r_w - r_e}$$

For a standard atmosphere it has been found that

$$r_w = 4r_e$$

Consequently,

$$k = 4/3$$

and the *effective radius of the earth* is

$$\begin{aligned} r_e' &= kr_e \\ &= 4/3r_e = 1.333r_e \end{aligned}$$

Actually, k varies over a range from 1.1 to about 1.6, the value varying directly as a function of the temperature and inversely with humidity.

REFERENCES

Books

Brainerd, J. G., Koehler, G., Reich, H. J., and Woodruff, I., *Ultra-High-Frequency Techniques*, Van Nostrand, 1942.

Everitt, W. L., *Communication Engineering*, McGraw-Hill.

Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill.

M.I.T. Staff, *Principles of Radar*, McGraw-Hill, 1946.

Periodicals

Gilliland, Kirby, Smith, and Reyner, "Characteristics of the Inosphere and Their Application to Radio Transmission," *Proc. IRE*, July 1937.

Mimno, Harry R., "The Physics of the Ionosphere," *Rev. Mod. Phys.*, Jan. 1937.

Smith, "Relation of Radio Sky Wave Transmission to Ionosphere Measurements," *Proc. IRE*, May 1939.

Superintendent of Documents, "Instructions for the Use of Basic Radio Propagation Predictions," *Gov. Printing Office.*

Superintendent of Documents, "Basic Radio Propagation Predictions," *Gov. Printing Office.*

APPENDIX 1
FUNDAMENTAL CONSTANTS

Electronic charge:

$$e = (1.600 \pm 0.002)\ 10^{-19}\ \text{coulomb}$$

Electronic mass:

$$m = (9.156 \pm 0.018)\ 10^{-31}\ \text{kg}$$

Ratio of electronic charge to mass:

$$e/m = (1.7571 \pm 0.0015)\ 10^{11}\ \text{coulomb/kg}$$

Dielectric constant of free space:

$$\epsilon_0 = 8.854 \times 10^{-12} \simeq \frac{1}{36\pi} \times 10^{-9}\ \text{farad/m}$$

Permeability of free space:

$$\mu_0 = 4\pi \times 10^{-7} \simeq 1.257 \times 10^{-6}\ \text{henry/m}$$

Velocity of light:

$$c = \sqrt{\frac{1}{\mu_0 \epsilon_0}} = 2.998 \times 10^{8}\ \text{m/sec}$$

APPENDIX II

SOLUTION OF THE CYLINDRICAL WAVEGUIDE

AT ANY ordinary point in space, Maxwell's equations in differential form are

$$\nabla \times \mathbf{E}' = -\mu \frac{\partial \mathbf{H}'}{\partial t} \tag{1}$$

$$\nabla \times \mathbf{H}' = \mathbf{J}' + \frac{\partial \mathbf{D}'}{\partial t} = \left(\sigma + \epsilon \frac{\partial}{\partial t}\right)\mathbf{E}' \tag{2}$$

In the usual case, both vectors may be assumed to be related to the direction of propagation z and to depend upon time in the manner specified below.

$$\mathbf{E}' = \mathbf{E}^0 e^{j\omega t - \gamma z} = \mathbf{E} e^{j\omega t} \quad \text{where} \quad \mathbf{E} = \mathbf{E}^0 e^{-\gamma z}$$
$$\mathbf{H}' = \mathbf{H}^0 e^{j\omega t - \gamma z} = \mathbf{H} e^{j\omega t} \quad \text{where} \quad \mathbf{H} = \mathbf{H}^0 e^{-\gamma z}$$

Substituting these relationships into equations (1) and (2) yields

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{3}$$
$$\nabla \times \mathbf{H} = (\sigma + j\omega\epsilon)\mathbf{E} \tag{4}$$

It can be shown that the curl of a vector $\mathbf{F}$ in cylindrical coordinates is given by

$$\nabla \times \mathbf{F} = \left(\frac{1}{r}\frac{\partial F_z}{\partial \theta} - \frac{\partial F_\theta}{\partial z}\right)\mathbf{i}_r + \left(\frac{\partial F_r}{\partial z} - \frac{\partial F_z}{\partial r}\right)\mathbf{i}_\theta + \left[\frac{1}{r}\frac{\partial}{\partial r}(rF_\theta) - \frac{1}{r}\frac{\partial F_r}{\partial \theta}\right]\mathbf{i}_z$$

Readers interested in the development of this relationship are referred to pages 50–52 of *Electromagnetic Theory*, by J. A. Stratton, McGraw-Hill, 1941.

Application of this relationship to Maxwell's equations will yield the following terms when components are equated.

$$\left.\begin{aligned}
&\frac{1}{r}\frac{\partial E_z}{\partial \theta} - \frac{\partial E_\theta}{\partial z} = -j\omega\mu H_r && \text{(a)}\\
&\frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r} = -j\omega\mu H_\theta && \text{(b)}\\
&\frac{1}{r}\frac{\partial}{\partial r}(rE_\theta) - \frac{1}{r}\frac{\partial E_r}{\partial \theta} = -j\omega\mu H_z && \text{(c)}
\end{aligned}\right\} \qquad (5)$$

$$\left.\begin{aligned}
&\frac{1}{r}\frac{\partial H_z}{\partial \theta} - \frac{\partial H_\theta}{\partial z} = (\sigma + j\omega\epsilon)E_r && \text{(a)}\\
&\frac{\partial H_r}{\partial z} - \frac{\partial H_z}{\partial r} = (\sigma + j\omega\epsilon)E_\theta && \text{(b)}\\
&\frac{1}{r}\frac{\partial}{\partial r}(rH_\theta) - \frac{1}{r}\frac{\partial H_r}{\partial \theta} = (\sigma + j\omega\epsilon)E_z && \text{(c)}
\end{aligned}\right\} \qquad (6)$$

In Article 7.15, pages 316–319, the derivation of the field equations for the cylindrical waveguide was carried through in terms of the *rectangular components* of the **E** and **H** vectors. However, the Laplacian operator in cylindrical coordinates was used. As long as the rectangular components of the field vectors are used, the original vector wave equation may be reduced to a scalar equation that must be satisfied by each of the rectangular components. This is true regardless of the coordinate system used for the Laplacian. However, it is not true when the *field vectors* are expressed in any other coordinate system, such as cylindrical coordinates.

It is evident that a difficulty has been introduced because our solution is carried through in terms of the rectangular field components, whereas we want the cylindrical components. Consequently, it becomes necessary to develop a method of converting the rectangular components obtained from a solution of the scalar wave equation in cylindrical coordinates to the desired cylindrical components.

It should be noted that it is not necessary to follow this particular method. It is possible to derive the vector wave equation in cylindrical coordinates from the $\nabla \times \nabla \times \mathbf{E}$ and $\nabla \times \nabla \times \mathbf{H}$ relationships. However, the procedure outlined in the preceding paragraphs is used in an effort to preserve a generalized viewpoint consistent with the method followed for the rectangular waveguide.

The equivalence between the various field components in the two coordinate systems may be established in the following manner.

Refer to the definitions of the coordinate systems given on page 316. We can write

$$E_x = E_r \cos\theta - E_\theta \sin\theta \qquad H_x = H_r \cos\theta - H_\theta \sin\theta$$
$$E_y = E_r \sin\theta + E_\theta \cos\theta \qquad H_y = H_r \sin\theta - H_\theta \cos\theta$$
$$E_z = E_z \qquad H_z = H_z$$

Note that the z component is common to both coordinate systems. Consequently, it would be advantageous to express the other cylindrical components in terms of the z components. This is readily done as follows. Solve equation (5b) for H_θ.

$$H_\theta = \frac{1}{-j\omega\mu}\left(\frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r}\right) = -\frac{1}{j\omega\mu}\left(\gamma E_r - \frac{\partial E_z}{\partial r}\right)$$

Hence,

$$H_\theta = \frac{\gamma E_r}{j\omega\mu} + \frac{1}{j\omega\mu}\frac{\partial E_z}{\partial r}$$

The derivative with respect to z is then

$$\frac{\partial H_\theta}{\partial z} = -\frac{\gamma^2}{j\omega\mu}E_r - \frac{\gamma}{j\omega\mu}\frac{\partial E_z}{\partial r}$$

Now substitute this into equation (6a).

$$\frac{1}{r}\frac{\partial H_z}{\partial \theta} + \frac{\gamma^2 E_r}{j\omega\mu} + \frac{\gamma}{j\omega\mu}\frac{\partial H_z}{\partial r} = (\sigma + j\omega\epsilon)E_r$$

Then, solving for E_r yields

$$E_r = \frac{1}{j\omega\mu(\sigma + j\omega\epsilon) - \gamma^2}\left(\gamma\frac{\partial E_z}{\partial r} + \frac{1}{r}\frac{\partial H_z}{\partial \theta}\right) \tag{7}$$

Application of this same method to the remaining components will give

$$E_\theta = \frac{1}{j\omega\mu(\sigma + j\omega\epsilon) - \gamma^2}\left(\frac{\gamma}{r}\frac{\partial E_z}{\partial \theta} - j\omega\mu\frac{\partial H_z}{\partial r}\right) \tag{8}$$

$$H_r = \frac{1}{j\omega\mu(\sigma + j\omega\epsilon) - \gamma^2}\left(\gamma\frac{\partial H_z}{\partial r} - \frac{j\omega\epsilon}{r}\frac{\partial E_z}{\partial \theta}\right) \tag{9}$$

$$H_\theta = \frac{1}{j\omega\mu(\sigma + j\omega\epsilon) - \gamma^2}\left(\frac{\gamma}{r}\frac{\partial H_z}{\partial \theta} + j\omega\epsilon\frac{\partial E_z}{\partial r}\right) \tag{10}$$

Equations (7) through (10) express the remaining cylindrical components in terms of the z components common to both coordinate systems.

It is evident then that the field equations defining operation in a cylindrical waveguide may be obtained by applying the scalar wave equation, as in Article 7.15, to the z component of both the **E** and **H** vectors. The remaining components are then evaluated from equations (7) through (10) in this Appendix.

INDEX