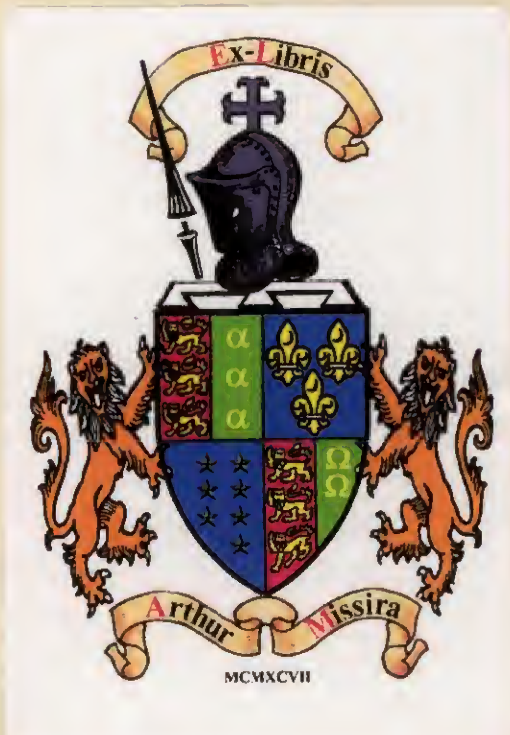
RADIO ENGINEERING

TERMAN

McGRAW-HILL BOOK COMPANY

David Powell



RADIO ENGINEERING

This book is produced in full compliance with the government's regulations for conserving paper and other essential materials.

RADIO ENGINEERING

BY

FREDERICK EMMONS TERMAN, Sc.D.

Professor of Electrical Engineering
Stanford University

SECOND EDITION
FIFTEENTH IMPRESSION

McGRAW-HILL BOOK COMPANY, INC.
NEW YORK AND LONDON
1937

PRINTED IN THE UNITED STATES OF AMERICA

THE MAPLE PRESS COMPANY, YORK, PA

PREFACE TO THE SECOND EDITION

The present edition of "Radio Engineering," while following the same method of presentation as the first edition, has been completely revised to take into account the rapid advances in radio communication during recent years. The extent of this revision is indicated by the fact that the entire book has been rewritten except for the first several chapters, and over half of the illustrations of the present edition are new.

The order and details of presentation have been modified in a number of instances in order more readily to include new developments and present trends. The fundamental principles of all types of electron tubes are now given in one unified chapter. The sections on power amplifiers have been expanded into a full chapter that represents the most complete treatment of this subject now available in book form. A brief chapter on television has also been added because of the increasing importance of this field. In order to take care of these and other additions without undue increase in length, the chapter on measurements has been omitted. It is believed that with the appearance of the author's book "Measurements in Radio Engineering" the need for such a chapter has been greatly lessened.

The present edition carries the analysis of many topics much farther than did the first edition. A complete treatment of the diode detector is presented, including the effects of the a-c–d-c ratio. The power-series method of analyzing distortion, modulation, etc., in tubes is fully covered, with the results boiled down to an equivalent circuit that greatly simplifies the application of this analysis to engineering problems. There is increased attention given also to the design and analysis of power amplifiers of all types, and of modulators. The chapter on wave propagation now includes sufficient material to permit calculation of ground-wave coverage, determination of optimum frequency for sky-wave transmission from ionospheric data, and the calculation of received field strength at ultra-high frequencies.

A certain amount of new material is presented here for the first time. This includes an accurate analysis of condenser-input filter systems, universal amplification curves of resistance- and transformer-coupled amplifiers, the analysis of the directional characteristics of radiation from a nonresonant wire in space, and the high-efficiency grid- and suppressor-grid-modulated amplifiers.

Problems are given at the end of each chapter. In compiling these an effort has been made to include at least one problem on every important topic in the text. This has been done even for such subjects as radio transmitters, for which it is difficult to formulate problems. As a result it will be possible for the teacher to supplement all text assignments with corresponding problem assignments that cannot be filled until an understanding of the corresponding parts of the text has been gained. In order to further this, the problems have in the main been so worded that they cannot be answered by copying from the text or by simple substitution in formulas, but rather call for an understanding of how to apply the principles given in the text. In using these problems, the instructor is expected to fit them to his own particular requirements. Thus the total number of problems available is greater than can be included in any ordinary radio course. Also, some of the problems are rather long and difficult, and an instructor may find it desirable to shorten them. Other problems are intentionally presented in a manner similar to the way the corresponding problem is encountered in practice. That is, the desired result is specified, and the student is asked to design apparatus and circuits that will give the required performance without any details such as number and types of tubes, etc., being given. It is believed such a comprehensive group of problems will suggest to an instructor the various possibilities that are present and will serve as the basis for problem sets suitable for any circumstance.

The author is indebted to a number of individuals and organizations for assistance rendered in the preparation of this edition. Mr. Harold A. Wheeler and the Hazeltine Corporation very generously made certain material available to the author before publication. Thanks are also due Bell Telephone Laboratories, RCA Radiotron Company, Inc., RCA Manufacturing Company, Inc., RCA Communications, Inc., Philco Radio & Television Corporation, The Crosley Radio Corporation, Jensen Radio Co., and the Stromberg-Carlson Telephone Company for photographs, circuit diagrams, and data on commercial equipment. Mention should also be made of Lowell Hollingsworth, F. Alton Everest, John Donahue, Robert Sink, J. S. Anderson, and John R. Woodyard, all students or former students at Stanford, who assisted in drawing figures and checking the manuscript and proof. Acknowledgment is made for suggestions, criticisms, and other help rendered by Dr. R. L. Freeman, formerly of Crosley Radio, and James M. Sharp, instructor at Stanford University.

FREDERICK EMMONS TERMAN.

STANFORD UNIVERSITY,
June, 1937.

PREFACE TO THE FIRST EDITION

The aim of "Radio Engineering" is to present a comprehensive engineering treatment of the more important vacuum-tube and radio phenomena. Electrical circuits and vacuum tubes behave according to exact laws, which in the main are simple and easily understood, and which can be used to predict the performance of radio circuits and radio apparatus with the same certainty and accuracy that the performance of other types of electrical equipment, such as transformers, motors, and transmission lines, is analyzed. It is this ability to reduce a problem to quantitative relations that predict with accuracy the performance to be expected or explain the results already obtained that represents a real mastery of the subject such as the radio engineer is expected to possess.

The principal prerequisite for undertaking the study of radio engineering is a good working knowledge of the fundamental concepts of alternating currents, such as reactance, impedance, power factor, phase angle, and vector representation. An elementary idea of complex quantity notation is also desirable but not absolutely essential. This means that radio work as outlined in this volume can be taken up in the senior year of the usual electrical engineering curriculum.

The order of presentation has been intentionally so arranged that the first part of the volume is devoted to the theory of tuned circuits and the fundamental properties of vacuum tubes and vacuum-tube applications, all of which are of importance and interest to every electrical engineer. The latter part then takes up more specialized radio topics, such as radio receivers and transmitters, wave propagation, antennas, and direction finding. This makes it possible where desired to arrange a two-semester course, the first term of which will be suitable for all electrical engineers, with the second term continuing for those with definite radio interests.

Particular care has been taken to avoid unnecessary equations in developing the analytical side of the various subjects taken up. It has been the author's experience that the usual student when first coming in contact with a new subject is confused by the presence of numerous equations and, in such circumstances, frequently fails to realize which relations are of real importance. By carrying on the reasoning in terms of words as far as possible, by judicious use of footnotes, and by skipping over purely routine mathematical manipulations, it has been possible to cut down the number of equations appearing in the text to the point where the important mathematical relations stand out by virtue of the

fact that they stand nearly alone, free of attention-diverting trivial equations. The result is that while "Radio Engineering" appears to be relatively free from mathematics, yet it actually carries the analysis much deeper than is customary. A typical illustration of this is the treatment of the transformer-coupled amplifier given in Chap. V. As far as the author is aware, this represents the only published analysis that can be used to predict the complete amplification characteristic of the transformer-coupled amplifier without an unreasonable amount of work and with engineering accuracy. At the same time it is almost devoid of mathematics as compared with the incomplete and often incorrect treatment ordinarily found. This result has been achieved by carrying the reasoning along in terms of physical concepts and words and by writing down an equation only when the equation itself is of importance.

A considerable quantity of original material is being published here for the first time. Notable instances of this are the analysis of the transformer-coupled amplifier mentioned above, the universal resonance curve, the Class A power amplifier formulas, the analysis of the Class B (linear) power amplifier, the analysis of regeneration resulting from a common plate impedance, the concept of the effective Q of the tuned amplifier, the analysis of the input admittance of amplifiers, the treatment of the voltage and current relations existing in the screen-grid tube, and the approximate analysis of rectifier-filter systems having a shunt condenser across the filter input.

The footnote references form an integral part of the text and have been carefully selected with a view toward helping the reader who desires more information on a particular subject than is given in this volume. No attempt has been made to compile complete bibliographies, the aim having been rather to cite a limited number of comprehensive articles that are really readable by the average student.

The author wishes to acknowledge the very helpful cooperation which has been received on all sides. Particular mention should be made of Philip G. Caldwell, the late Nathaniel R. Morgan, Paul F. Byrne, Dr. Horace E. Overacker, William R. Triplett, Harry Engwicht, W. G. Wagener, and D. A. Murray, all former students at Stanford University, who assisted in drawing the figures and checking the manuscript and proof. The author is also greatly indebted to the Bell Telephone Laboratories, the American Telephone and Telegraph Company, the General Radio Company, the De Forest Radio Company, the General Electric Company, RCA Radiotron, Inc., and the Stromberg-Carlson Telephone Manufacturing Company for supplying copy for certain of the illustrations.

STANFORD UNIVERSITY, CALIFORNIA, FREDERICK EMMONS TERMAN.
August, 1932.

CONTENTS

RADIO ENGINEERING

CHAPTER I

THE ELEMENTS OF A SYSTEM OF RADIO COMMUNICATION

1. Radio Waves.—Electrical energy that has escaped into free space exists in the form of electromagnetic waves. These waves, which are commonly called radio waves, travel with the velocity of light and consist of magnetic and electrostatic fields at right angles to each other and also at right angles to the direction of travel. If these electrostatic and magnetic fluxes could actually be seen, the wave would have the appearance indicated in Fig. 1. One-half of the electrical energy contained

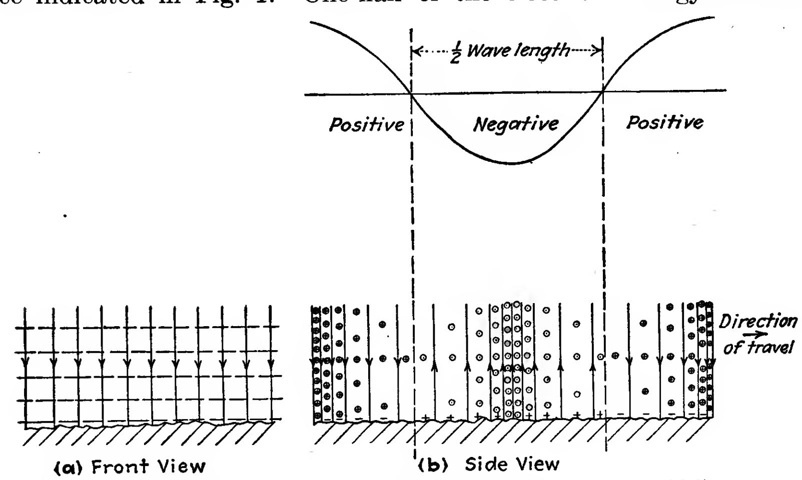


Fig. 1.—Front and side views of a vertically polarized wave. The solid lines represent electrostatic flux, while the dotted lines and the circles indicate magnetic flux.

in the wave exists in the form of electrostatic energy, while the remaining half is in the form of magnetic energy.

The essential properties of a radio wave are the frequency, intensity, direction of travel, and plane of polarization. The radio waves produced by an alternating current will vary in intensity with the frequency of the current and will therefore be alternately positive and negative as shown at *b* in Fig. 1. The distance occupied by one complete cycle of such an alternating wave is equal to the velocity of the wave divided by the number of cycles that are sent out each second and is called the wave length.

The relation between wave length λ in meters and frequency f in cycles per second is therefore

$$\lambda = \frac{300,000,000}{f} \tag{1}$$

The quantity 300,000,000 is the velocity of light in meters per second. The frequency is ordinarily expressed in kilocycles, abbreviated kc; or in megacycles, abbreviated mc. A low-frequency wave is seen from Eq. (1) to have a long wave length, while a high frequency corresponds to a short wave length.

The strength of a radio wave is measured in terms of the voltage stress produced in space by the electrostatic field of the wave and is usually expressed in microvolts stress per meter. Since the actual stress produced at any point by an alternating wave varies sinusoidally from instant to instant, it is customary to consider the intensity of such a wave to be the effective value of the stress, which is 0.707 times the maximum stress in the atmosphere during the cycle. The strength of the wave measured in terms of microvolts per meter of stress in space is exactly the same voltage that the magnetic flux of the wave induces in a conductor 1 meter long when sweeping across this conductor with the velocity of light. Thus the strength of a wave is not only the dielectric stress produced in space by the electrostatic field, but it also represents the voltage that the magnetic field of the wave will induce in cutting across a conductor. In fact, the voltage stress produced by the wave can be considered as resulting from the movement of the magnetic flux of the same wave.

The minimum field strength required to give satisfactory reception of the radio wave varies with the amount of interference that is present. Under the most favorable conditions it is possible to obtain intelligible signals from waves having a strength as low as 0.1 μv per meter, but ordinarily interfering waves generated by both man-made and natural sources drown out such weak radio signals and make much greater field strengths necessary. Thus experience has shown that in rural areas a field strength in the order of 100 μv per meter is required to give what the listener considers satisfactory service from a broadcast station, while in urban locations, where the man-made interference is much greater, a field strength of 5000 to 30,000 μv per meter is needed to insure good reception at all times.

A plane parallel to the mutually perpendicular lines of electrostatic and electromagnetic flux is called the wave front. The wave always travels in a direction at right angles to the wave front, but whether it goes forward or backward depends upon the relative direction of the lines of electromagnetic and electrostatic flux. If the direction of either the

magnetic or electrostatic flux is reversed, the direction of travel is reversed, but reversing both sets of flux has no effect.

The direction of the electrostatic lines of flux is called the direction of polarization of the wave. If the electrostatic flux lines are vertical, as shown in Fig. 1, the wave is vertically polarized; when the electrostatic flux lines are horizontal and the electromagnetic flux lines are vertical, the wave is horizontally polarized.

Propagation of Radio Waves of Different Frequencies.—As radio waves travel away from their point of origin, they become attenuated as a result of spreading out and because of energy lost in travel. The amount of this attenuation depends upon the frequency of the wave, the time of day, the season of the year, and the character of the earth's surface.

Waves having frequencies below about 100 kc are called low-frequency waves, and travel with an attenuation that is small and relatively independent of time of day and season. These frequencies are therefore well suited for carrying on continuous radio communication over distances as great as 5000 miles.

Frequencies ranging from 100 to 1500 kc are referred to as medium radio frequencies. The distinguishing characteristic of these waves is high received energy at night, particularly in the winter, and high attenuation during the day. This effect becomes more pronounced as the frequency is increased toward 1500 kc and is well illustrated by the characteristics of broadcast waves. Medium-frequency waves are suitable for covering distances up to several thousand miles at night but only a few hundred miles in the daytime.

As the frequency is increased from 1500 toward 6000 kc, the attenuation, particularly during the day, becomes less, and these waves, which are said to be of medium-high frequency, are therefore suitable for carrying on communication over distances in the order of several thousand miles. The attenuation is greater during the day than at night, and greater in summer than in winter, but is sufficiently small even under unfavorable conditions to permit communication over distances of one or more thousands of miles.

Waves of frequencies ranging from 6000 to about 30,000 kc are called high-frequency waves. They are capable of traveling great distances with small attenuation, but whether or not they reach a particular destination depends upon the conditions existing in the ionized regions in the upper atmosphere. In general, the frequency best suited for reaching a very distant receiving point is highest on a summer day, somewhat lower on a winter day or summer night, and lowest during a winter night. In order to maintain reasonably continuous communication over great distances using high frequencies, it is necessary to change the frequency of transmission as conditions warrant.

Radio waves having frequencies less than about 12 kc have not been found useful for commercial radio communication because these waves require such large radiators for efficient radiation as to be impracticable from an economic point of view. Waves having frequencies higher than about 30,000 kc are likewise of limited use because they travel along straight lines and are not reflected back to earth by the ionized region in the upper atmosphere. They are therefore of use only over distances so short that the earth's curvature permits a substantially straight-line path between transmitting and receiving points.

TABLE I.—CLASSIFICATION OF RADIO WAVES

Class	Frequency range, kilocycles	Wave-length range, meters	Outstanding characteristics	Principal uses
Low frequency....	Below 100	Over 3,000	Low attenuation at all times of day and of year	Long-distance transoceanic service requiring continuous operation
Medium frequency	100 to 1,500	3,000 to 200	Attenuation low at night and high in the daytime; greater in summer than winter	Range 100 to 500 kc used for marine communication, airplane radio, direction finding, etc. Range 550 to 1500 kc employed for broadcasting
Medium high frequency.	1,500 to 6,000	200 to 50	Attenuation low at night and moderate in the daytime	Moderate-distance communication of all types
High frequency...	6,000 to 30,000	50 to 10	Transmission depends solely upon the ionization in the upper atmosphere, and so varies greatly with the time of day and season. Attenuation extremely small under favorable conditions	Long-distance communication of all kinds; airplane radio
Very high frequency.	Above 30,000	Below 10	Waves travel in straight lines and are not reflected by ionized layers, so can only travel between points in sight of each other	Short-distance communication, television, two-way police radio, portable equipment, airplane landing beacons

As a result of the different characteristics of propagation possessed by radio waves of different frequencies, each particular range of frequencies is best adapted for a particular type of communication service. The outstanding properties of the different classes of radio waves are tabulated in Table I, as well as the uses to which each class has been found best suited.

2. Radiation of Electrical Energy.—Every electrical circuit carrying alternating current radiates a certain amount of electrical energy in the form of electromagnetic waves, but the amount of energy thus radiated is extremely small unless all the dimensions of the circuit approach the order of magnitude of a wave length. Thus a power line carrying 60-cycle current with 20-ft. spacing between conductors will radiate practically no energy because a wave length at 60 cycles is more than 3000 miles and 20 ft. is negligible in comparison. On the other hand, a coil 20 ft. in diameter and carrying a 2000-kc current will radiate a considerable amount of energy because 20 ft. is comparable with the 150-meter wave length of the radio wave. The common radio antenna consisting of a vertical wire with a flat-top structure as shown in Fig. 2 is essentially a condenser in which one plate is the ground while the other plate is the flat top. Such an arrangement will be a good radiator of electrical energy when the ratio of height to wave length is appreciable, *i.e.*, at least 1:100, and preferably 1:10. Similarly a coil will be a good radiator of electrical energy provided the size of the coil is sufficiently great. The usual loop antenna consists of a coil and will be an efficient radiator to the extent that the ratio of loop diameter to wave length is appreciable.

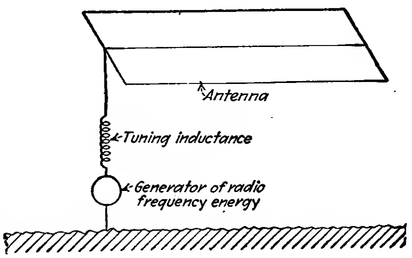


FIG. 2.—A simple system for producing radio waves, consisting of flat-top antenna, tuning inductance to bring antenna circuit into resonance at the frequency of the waves to be radiated, and a generator of radio-frequency energy.

It is apparent from these considerations that the size of radiator required is inversely proportional to the frequency. High-frequency waves can therefore be produced by a small radiator, while low-frequency waves require a high antenna system for effective radiation. The practical result of this fact is that the antennas of low-frequency transmitting stations are sometimes suspended from towers over 500 ft. high and yet are less efficient radiators than an antenna of one-tenth this height operating on a very high radio frequency.

Every radiator has directional characteristics as a result of which it sends out stronger waves in certain directions than in others. Thus, while a vertical wire radiates the same amount of energy in directions

that are perpendicular to the wire, the radiation in a vertical plane varies from a maximum in a horizontal direction to zero in a vertical direction, as shown in Fig. 3. Directional characteristics of antennas are taken advantage of to concentrate the radiation toward the point to which it is desired to transmit.

The amount of energy sent out from any radiating system is proportional to the square of the radio-frequency current that flows in the radiator. Since all the common sources of radio-frequency energy are relatively low-voltage high-current sources, it is necessary that the radiating system offer a relatively low impedance to the radio-frequency energy that is to be transmitted. This is accomplished by tuning the antenna circuit to resonance with the frequency to be radiated, which makes the impedance of the antenna circuit low and enables a relatively small applied voltage to produce a very large antenna current and hence a high radiated energy. This is the only reason for tuning the transmitting antenna, as the mere tuning of the radiating systems to the frequency being transmitted does not increase the radiated power per ampere of current. The tuning is accomplished by inserting an inductance or a condenser in series with the antenna, as circumstances require. Thus in the flat-top antenna of Fig. 2 the antenna has a capacity reactance and so is tuned by the use of the inductance coil shown in the figure.

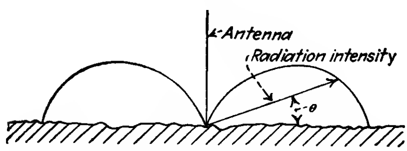


FIG. 3.—Directional characteristics in a vertical plane of radiation from an antenna consisting of a vertical wire. The length of the radius vector from the base of the antenna represents the relative intensity of wave radiated in the direction of the vector.

3. Generation and Control of Radio-frequency Power.—The radio-frequency power required by the radio transmitter is practically always obtained from a vacuum-tube oscillator. Vacuum-tube oscillators are capable of converting direct-current power into alternating-current energy of any desired frequency up to 300,000,000 cycles or higher. Over the range of frequencies used in long-distance radio communication, *i.e.*, 12 to 30,000 kc, the power that can be obtained from vacuum-tube oscillators is in the order of tens to hundreds of kilowatts, and the efficiency with which the direct-current power is transformed into alternating-current energy is in the neighborhood of 50 per cent or higher.

A number of other methods of obtaining radio-frequency energy have been used at one time or another during the history of radio. Among these are the high-frequency alternator, the Poulsen arc, the frequency multiplier, and the oscillatory spark discharge. The high-frequency alternator is a special high-speed inductor-type alternator with many poles. Such alternators can generate several hundreds of kilowatts with reasonable efficiency when operating at frequencies of

50,000 cycles or less. A number of high-frequency alternators are now in commercial use, although it is improbable that any more will ever be built.[1] The frequency multiplier utilizes a moderately high frequency alternator from which the desired radio frequency is obtained by the use of magnetic-type harmonic generators. In this way it is possible with an alternator giving a frequency of 5000 cycles to produce considerable quantities of power at frequencies of from 20,000 to 40,000 cycles. This type of arrangement at one time had a very prominent place in radio but has now practically disappeared.[2]

The Poulsen arc takes advantage of the negative-resistance characteristic of an electric arc to convert direct-current power into radio-frequency energy. The arc as ordinarily employed takes place between carbon and copper electrodes in an atmosphere of hydrocarbon vapor and with a magnetic field at right angles to the axis of the arc. Such an arrangement, when properly designed, will generate large quantities of radio-frequency energy at a fair efficiency. The Poulsen arc operates most efficiently at frequencies below several hundred thousand cycles, but it will function after a fashion up to frequencies approaching 2000 kc. For the frequencies to which it is adapted the arc is a very simple and rugged generator of radio-frequency energy, and hundreds are still in regular use. It is, however, being rapidly replaced by the vacuum-tube oscillator because of the latter's flexibility, frequency stability, and freedom from harmonics.[3]

The oscillatory spark discharge was the earliest and for many years the only method known for the generation of radio-frequency power. In this type of transmitter a condenser is charged to a high potential, which then breaks down a spark gap, permitting an oscillatory discharge through an inductance. This process is repeated about 1000 times each second. The spark transmitter thus radiates a series of wave trains, each of which is a damped sinusoidal oscillation. This method is capable of generating large quantities of radio-frequency energy with

[1] A description of the widely used Alexanderson alternator is to be found in an article by Ernst F. W. Alexanderson, Trans-oceanic Radio Communication, *Proc. I.R.E.*, vol. 8, p. 263, August, 1920. For information on this as well as other types of alternators see G. G. Blake, "History of Radio Telegraphy and Telephony," pp. 230–232, Chapman & Hall, Ltd.

[2] Extensive discussions of frequency multipliers are to be found in practically every book on radio communication written prior to about 1920.

[3] For further information regarding the Poulsen arc, the reader is referred to the following articles: Leonard F. Fuller, The Design of Poulsen Arc Converters for Radio Telegraphy, *Proc. I.R.E.*, vol. 7, p. 449, October, 1919; P. O. Pedersen, On the Poulsen Arc and Its Theory, *Proc. I.R.E.*, vol. 5, p. 255, August, 1917; Some Improvements in the Poulsen Arc, *Proc. I.R.E.*, vol. 9, p. 434, October, 1921, and vol. 11, p. 155, April, 1923.

good efficiency but is in disfavor because the radiated waves are not simple sine waves but rather waves of a number of frequencies superimposed on each other. The result is excessive interference with radio signals being transmitted on slightly different frequencies.[1]

Modulation.—The transmission of information by radio waves requires that some means be employed to control the radio waves by the desired

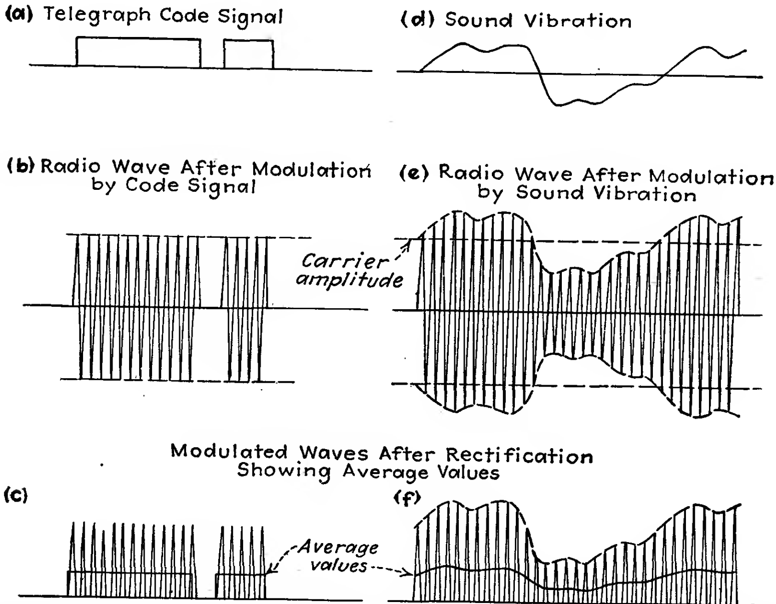


FIG. 4.—Diagram showing how a signal may be transmitted by modulating the amplitude of a radio wave, and how the original signal may be recovered from the modulated wave by rectification.

intelligence. In radio telegraphy this control is obtained by turning the transmitter on and off in accordance with the dots and dashes of the telegraph code, as illustrated in Fig. 4. In radio telephony the transmission is accomplished by varying the amplitude of the radio-frequency wave in accordance with the pressure of the sound wave being transmitted. Thus the sound wave shown at *d* in Fig. 4 would be transmitted from a radio-telephone station by causing the amplitude of the radiated wave to vary as shown at *e*. In the transmission of pictures by radio a similar method is employed, in which the amplitude of the wave radiated

[1] An extensive art has been developed in connection with the spark generator of radio-frequency energy, which, though obsolete as far as radio communication is concerned, is still of value in other applications. Interested readers will find an excellent treatment of the spark generator in J. H. Morecroft, "Principles of Radio Communication," 2d ed., Chap. V., John Wiley & Sons, Inc.

at any time is made proportional to the light intensity of the part of the picture that is being transmitted at that instant.

When the amplitude of the alternating-current wave is varied from time to time, the wave is said to be modulated. Thus the wave radiated from a radio-telephone station is modulated by the voice or sound wave, while during the transmission of a picture the modulation is in accordance with the light intensities of different portions of the picture, and in the case of radio telegraphy the modulation is by the telegraph code. Except in the case of telegraphy, the modulation of the radio-frequency wave is usually accomplished by means of vacuum tubes that control the amplitude of the generated or radiated high-frequency energy in accordance with the intelligence that is to be transmitted.

4. Reception of Radio Signals.—In the reception of radio signals it is first necessary to abstract energy from the radio waves passing the receiving point. After this has been done, the radio receiver must first separate the desired signal from other signals that may be present, and then reproduce the original intelligence from the radio waves. In addition, arrangements are ordinarily provided for amplification of the received energy so that the output of the radio receiver can be greater than the energy abstracted from the wave.

Any antenna system capable of radiating electrical energy is also able to abstract energy from a passing radio wave because the electromagnetic flux of the wave in cutting across the antenna conductors induces a voltage that varies with time in exactly the same way as the current flowing in the antenna radiating the wave. The voltage induced in an antenna is equal to the product of the effective antenna height and the strength of wave, and the resulting current flowing in the antenna is the current that is produced by this induced voltage acting against the impedance of the circuit. The energy represented by the induced current flowing in the antenna system is abstracted from the passing wave and will be greatest when the impedance of the antenna system has been reduced to a minimum by making the antenna circuit resonant to the frequency of the wave to be received.

The characteristics of an antenna when used for receiving radio signals are the same as in sending. Thus an efficient transmitting antenna is equally efficient when used for the reception of radio signals, and in general the size of receiving antenna will be proportional to the wave length being received, just as is the case in transmitting. The directional characteristics of an antenna are the same for reception as transmission; that is, if an antenna radiates most of the energy delivered to it in one direction, then that antenna will have a larger voltage induced in it by waves coming from the direction in which radiation is large than by waves coming from other directions. Similar types of aerial systems are used

for both radiating and receiving electromagnetic waves, the only difference being that receiving antennas can be less expensive and smaller than transmitting antennas because amplification at the receiver can readily make up for a less efficient antenna.

Since every wave passing the receiving antenna induces its own voltage in the antenna conductor, it is necessary that the receiving equipment be capable of separating the desired signal from the unwanted signals that are also inducing voltages in the antenna. This separation is made on the basis of the difference in frequency between transmitting stations and is carried out by the use of resonant circuits which can be made to discriminate very strongly in favor of a particular frequency. It has already been pointed out that, by making the antenna circuit resonant to a particular frequency, the energy abstracted from radio waves of that frequency will be much greater than the energy from waves of other frequencies; this alone gives a certain amount of separation between signals. Still greater selective action can be obtained by the use of additional suitably adjusted resonant circuits located somewhere in the receiver in such a way as to reject all but the desired signal. The ability to discriminate between radio waves of different frequencies is called *selectivity* and the process of adjusting circuits to resonance with the frequency of a desired signal is spoken of as *tuning*.

Detection.—The process by which the signal being transmitted is reproduced from the radio-frequency currents present at the receiver is called detection. Where the intelligence is transmitted by varying the amplitude of the radiated wave, detection is accomplished by rectifying the radio-frequency currents. The rectified current thus produced varies in accordance with the signal originally modulated on the wave radiated at the transmitter and so reproduces the desired signal. Thus when the modulated wave shown at *e* of Fig. 4 is rectified, the resulting current is shown at *f* and is seen to have an average value that varies in accordance with the amplitude of the original signal. In the transmission of code signals by radio, the rectified current reproduces the dots and dashes of the telegraph code as shown at Fig. 4*c* and could be used to operate a telegraph sounder. When it is desired to receive the telegraph signals directly on a telephone receiver, it is necessary to break up the dots and dashes at an audible rate in order to give a note that can be heard, since otherwise the telephone receiver would give forth a succession of unintelligible clicks.

Although intelligible radio signals have been received from stations thousands of miles distant, using only the energy abstracted from the radio wave by the receiving antenna, much more satisfactory reception can be obtained if the received energy is amplified. This amplification may be applied to the radio-frequency currents before detection, in

which case it is called radio-frequency amplification, or it may be applied to the rectified currents after detection, which is called audio-frequency amplification. The use of amplification makes possible the satisfactory reception of signals from waves that would otherwise be too weak to give an audible response. It also permits the strength of the signal as heard in the telephone receiver or any other indicating device to be raised to any desired volume, permitting radio reception in noisy locations, such as on airplanes, and making possible the use of loud-speakers.

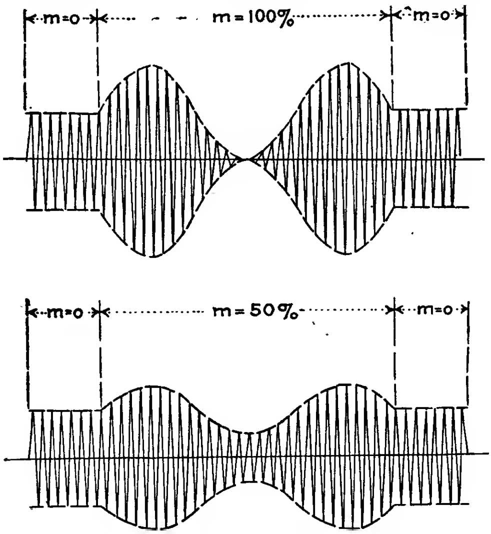


FIG. 5.—Waves having different degrees of simple sine-wave modulation.

The only satisfactory method of amplifying radio signals that has been discovered is by the use of vacuum tubes, and before such tubes were discovered radio reception had available only the energy abstracted from the radio wave by the receiving antenna. As a result of the small amplitude of this energy, the signals were always weak, and radio reception from other than local stations was possible only in very quiet places.

5. Nature of a Modulated Wave.—The modulated wave that is sent out by a radio station represents an oscillation of varying amplitude and so consists of a number of waves of different frequencies superimposed upon each other. The actual nature of a modulated wave can be deduced by writing down the equation of the wave and making a mathematical analysis of the result. Thus, in the case of the simple sine-wave modulations shown in Fig. 5, the amplitude of the radio-frequency oscillation is given by $E = E_0 + mE_0 \sin 2\pi f_s t$, in which E_0 represents the average amplitude, f_s the frequency at which the amplitude is varied, and m the ratio of amplitude variation from the average to the average ampli-

tude, which is called the degree of modulation. Waves with several degrees of modulation are shown in Fig. 5. The equation of these modulated waves can be written as

$$e = E_0(1 + m \sin 2\pi f_s t) \sin 2\pi f t \tag{2}$$

in which f is the frequency of the radio oscillation. Multiplying out the right-hand side of Eq. (2) gives

$$e = E_0 \sin 2\pi f t + mE_0 \sin 2\pi f_s t \sin 2\pi f t$$

By expanding the last term into functions of the sum and difference angles by the usual trigonometric formula, the equation of a wave with simple sine-wave modulation is

$$e = E_0 \sin 2\pi f t + \frac{mE_0}{2} \cos 2\pi(f - f_s)t - \frac{mE_0}{2} \cos 2\pi(f + f_s)t \tag{3}$$

Equation (3) shows that the wave with sine-wave modulation consists of three separate waves. The first of these is represented by the term $E_0 \sin 2\pi f t$ and is called the carrier. Its amplitude is independent of the presence or absence of modulation and is equal to the average amplitude of the wave, which is independent of the degree of modulation. The two other components are alike as far as magnitude is concerned, but the frequency of one of them is less than that of the carrier frequency by an amount equal to the modulation frequency, while the frequency of the second is more than that of the carrier by the same amount. These two components are called the side-band frequencies and carry the intelligence that is being transmitted by the modulated wave. The frequency by which the side bands differ from the carrier frequency represents the modulation frequency, while the amplitude of the side-band components compared with the amplitude of the carrier determines the degree of modulation, *i.e.*, the size of the amplitude variations that are impressed upon the radiated wave.

When the modulation is more complex than the simple sine-wave-amplitude variation of Fig. 5, the effect is to introduce additional side-band components. The carrier wave is always the same, irrespective of the character of the modulation, and represents the average amplitude of the wave, but there is a pair of side-band frequencies for each frequency component in the modulation. Thus, if the wave of a radio-telephone transmitter is modulated by a complex sound wave containing pitches of 1000 and 1500 cycles, the modulated wave will contain one pair of 1000-cycle side-band components and one pair of 1500-cycle side-band components. The amplitude of any side-band component is always one-half of the amplitude of that particular frequency component which is contained in the modulation envelope.

Significance of the Side Bands.—The carrier and side-band frequencies are not a mathematical fiction, but have a real existence, as is evidenced by the fact that the various frequency components of a modulated wave can be separated from each other by suitable filter circuits. The side-band frequencies can be considered as being generated as a result of varying the amplitude of the wave. They are present only when the amplitude is being varied, and their magnitude and frequency are determined by the character of the modulation. The carrier frequency, on the other hand, is independent of the modulation, being the same even when no modulation is present.

The intelligence transmitted by the modulated wave is carried by the side-band components and not by the carrier, *i.e.*, the intelligence is conveyed by the variations in the amplitude of the wave and not by the average amplitude. It is therefore desirable to put as much power into the side-band frequencies as is possible, which is equivalent to saying that the wave amplitude should be varied through the widest possible range. When the amplitude is carried clear to zero during the modulation cycle, the modulation is at a maximum, or 100 per cent, and the side bands contain the maximum amount of power possible. With sine-wave modulation such as shown in Fig. 5 this maximum side-band power is one-half of the carrier power. With degrees of modulation less than 100 per cent the side bands will contain correspondingly less power.

It is apparent that the transmission of intelligence requires the use of a band of frequencies rather than a single frequency. In speech and music there are important frequency components as high as 5000 cycles, so that speech and music modulated upon a wave will produce side-band components extending as far as 5000 cycles on each side of the carrier frequency. A radio-telephone station therefore utilizes a frequency band about 10,000 cycles wide in transmitting high-quality signals. If this entire band is not transmitted equally well through space and by the circuits through which the modulated wave currents must pass, then the side-band frequency components that are discriminated against will not be reproduced in the receiving equipment with proper amplitude. With telegraph signals the required side band is relatively narrow because the amplitude of the signals is varied only a few times a second, but a definite frequency band is still required. If some of the side-band components of the code signal are not transmitted, the received dots and dashes run together and may become indistinguishable.

CHAPTER II

CIRCUIT CONSTANTS

6. Inductance.—Whenever a current flows in an electrical circuit, there is produced magnetic flux that links with (*i.e.*, encircles) the current. The amount of such flux actually present with a given current is measured in terms of a property of the circuit called the inductance and depends upon the arrangement of the circuit and the presence or absence of magnetic substances.

Inductance can be defined as the flux linkages per ampere of current producing the flux; that is

$$\text{Inductance } L \text{ in henries} = \frac{\text{flux linkages}}{\text{current producing flux}} 10^{-8} \qquad (4)$$

A flux linkage represents one flux line encircling the circuit current once. Thus in Fig. 6 flux line *aa* contributes eight flux linkages toward the coil inductance because it circles the current flowing in the coil eight times. On the other hand, flux line *b* of the same coil contributes only one-half of a flux linkage toward the coil inductance because this particular line encircles only one-half of the coil current.

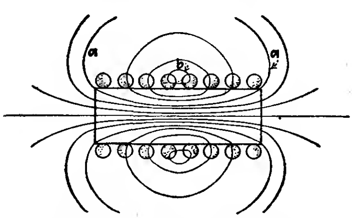


FIG. 6.—Flux and current distribution in typical single-layer air-cored inductance coil. The current density is indicated by the depth of shading.

The inductance of an electrical circuit is computed by assuming a convenient current flowing in the circuit, determining the magnetic flux produced by the current, and then counting up the total number of flux linkages present in the circuit. The inductance in henries is this total number of flux linkages multiplied by 10^{-8} and divided by the circuit current.

The inductance of a circuit is the measure of a number of electrical properties. When the current varies, as, for example, in the case of an alternating current, the amount of flux linking with the circuit also varies, and in doing so induces a voltage in the circuit just as magnetic flux always does when cutting a conductor. The magnitude of this induced voltage can be expressed in terms of the circuit inductance and rate of change of current.

$$\text{Induced voltage in circuit caused by change in current} = -L\frac{di}{dt} \qquad (5)$$

A positive voltage in this equation is a voltage acting in the direction of the current. The voltage that must be applied to the circuit to overcome the induced potential is equal and opposite to that given by Eq. (5). It will be observed that the voltage induced in the circuit by an increase of current is always in a direction that tends to reduce the current, while a decreasing current gives a negative di/dt and produces a positive induced voltage that tends to keep the current flowing. For the particular case where the current is alternating, $di/dt = j\omega I$, and substitution in Eq. (5) shows that the applied voltage required to overcome the induced voltage is the familiar quantity $j\omega LI$.

Flux linkages can be created only by the expenditure of energy, and these linkages represent electrical energy stored in the form of magnetic flux. The amount of such storage depends upon the circuit inductance and current according to the relation[1]

$$\text{Energy in joules stored in magnetic field} = \tfrac{1}{2}LI^2 \qquad (6)$$

Inductance Coils with Non-magnetic Cores.—Inductance coils intended for use at radio frequencies usually have non-magnetic cores. This is because the energy losses in magnetic materials are very great even at broadcast and lower frequencies unless the magnetic material is finely subdivided (see Sec. 12). Inductance coils with non-magnetic cores are also used at audio frequencies when it is necessary to avoid effects introduced by the variation of the permeability of magnetic materials with changing flux density.

Skilled mathematicians have derived formulas that give the inductance of all the commonly used types of coils with non-magnetic cores in terms of the coil dimensions. These formulas are usually both complicated and hard to derive because of the difficulty encountered in calculating the magnetic flux produced by a current flowing in the coil. In order to make such formulas of practical value, they are always simplified by the use of coefficients. Thus the inductance of a single-layer solenoid, such as shown in Fig. 6, is given by the formula

$$\text{Inductance in microhenries} = N^2dF \qquad (7)$$

where

N = number of turns

d = diameter of coil measured to center of wire

F = constant that depends only upon the ratio of length to diameter and is given in Fig. 7.

[1] Equations (5) and (6) are direct consequences of Eq. (4) and are derived in most textbooks on alternating currents. For example, see V. Karapetoff, "The Magnetic Circuit," Chap. X, McGraw-Hill Book Company, Inc.

The quantity F depends in a complicated way upon the ratio of coil length to diameter, but, once this relationship has been determined, the value of F for different ratios can be computed once for all and presented in a curve such as Fig. 7 or by a table.

Formulas giving the inductance of all the commonly used types of coils having non-magnetic cores are given in Appendix A. By the aid

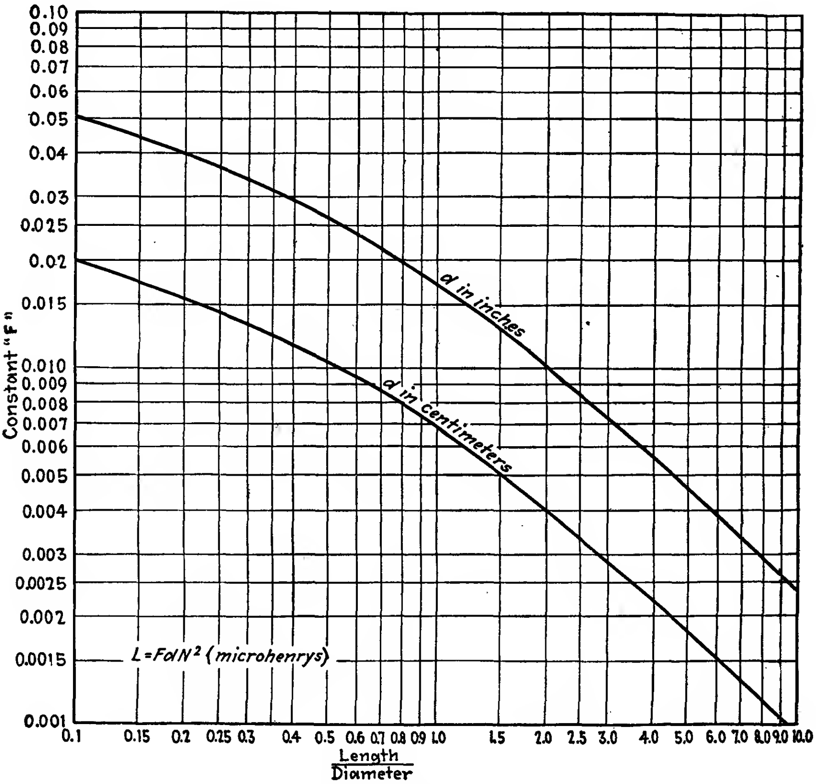


FIG. 7.—Values of constant F for use in Eq. (7), to obtain the inductance of single-layer solenoid.

of these formulas it is possible to compute the inductance of practically any type of coil with as great accuracy as the geometrical dimensions are ordinarily known. Guessing at the number of turns and the coil dimensions to obtain a desired inductance is therefore never necessary.

The inductance of all coils with non-magnetic cores is proportional to the square of the number of turns if the dimensions such as length, diameter, depth of winding, etc., are kept constant as the number of turns is altered. The reason for this behavior lies in the fact that, if the coil dimensions are kept constant, the amount of magnetic flux produced

by a given coil current and the number of times each flux line links with the coil current are both proportional to the number of turns.

The inductance of all coils having the same number of turns and the same shape is always proportional to the size (*i.e.*, to a linear dimension, such as length or radius) of the coil. Thus, if two coils have the same number of turns, but one is twice as big as the other in every dimension (such as diameter, length, width, depth of winding, etc.), then the larger coil will have twice the inductance of the smaller one. This rule can be verified by examining the inductance formulas in Appendix A; it results from the fact that the cross section of the flux paths is proportional to the square of the linear dimension of the coil, while the length of these paths varies directly as the linear dimension.

Inductance Coils with Magnetic Cores.—Magnetic cores greatly increase the flux produced by a given magnetomotive force (*i.e.*, a given number of ampere turns), and so make it possible to obtain a large inductance in a small volume and with the use of only a small amount of material. Because of this, inductance coils with magnetic cores are used in preference to coils with non-magnetic cores wherever this is possible. There are, however, two important factors limiting the usefulness of magnetic materials in coils intended for communication work. In the first place the permeability of all magnetic materials varies with the flux density and the previous magnetic history so that the inductance of a coil with a magnetic core depends somewhat upon the current flowing through the winding; furthermore, when this current is alternating, the inductance will be different at different parts of the cycle, which causes the production of harmonics. In the second place the eddy-current losses increase so rapidly with frequency as to make magnetically cored coils useless even at the lower radio frequencies unless special methods of subdividing the core are employed. Even with the greatest subdivision of the core material that it has been possible to obtain, the losses become excessive at frequencies of the order of 1000 kc and more (see Sec. 12 for further discussion).

It is relatively easy to compute the inductance of coils with magnetic cores with at least fair accuracy because the high permeability of magnetic materials restricts practically the entire flux to the magnetic core. This results in simple flux paths that can be readily handled in computations when the magnetic characteristics of the core material are known. The inductance of ordinary types of coils with magnetic cores is given to a good approximation by the formula[1]

$$L = 1.25N^2P \cdot 10^{-8} \tag{8}$$

[1] Equation (8) is derived as follows: The flux produced by a current I flowing in a coil of N turns with a core of permeance P is $1.257INP$ (*i.e.*, magnetomotive force

where

L = inductance in henries
N = number of turns
P = permeance of the magnetic circuit of coil, assuming permeability of air as unity.

This equation can be used in the calculation of inductance coils having a short air gap in the magnetic circuit provided the permeance P in the equation is interpreted to mean the permeance of the complete magnetic circuit, which in this case consists of the air gap in series with magnetic material.

The inductance of coils with magnetic cores is proportional to the square of the number of turns if the dimensions of the coil are kept constant, and is proportional to the linear dimensions if the number of turns and proportions are the same. These two relations are identical with those that exist in coils with non-magnetic cores and arise from the same reasons.

Incremental Permeability.—The behavior of coils having magnetic cores may become very complicated when the current through the winding contains several components of different frequencies. The most important example of this character is when an alternating current is superimposed upon a direct current. In this case the effective inductance offered to the alternating-current component depends upon the magnitudes of both currents and upon the previous magnetic history of the core. When a core that has been thoroughly demagnetized is first magnetized, the relation between current in the winding and core flux is the usual B-H curve shown at OA in Fig. 8. If the current is then successively reduced to zero, reversed, brought back to zero, reversed to the original direction, etc., the flux goes through the familiar hysteresis loop shown in Fig. 8. A direct current flowing through the coil winding brings the magnetic state of the core to some point on the hysteresis loop, such as B in Fig. 8, and it is to be observed that the flux with a given value of direct current depends not only upon the magnitude of current but also upon which side of the hysteresis loop one is on, and upon the width of the loop, both of which are determined by the previous magnetic history.

When an alternating current is now superimposed on this direct current, the result is to cause the flux in the iron core to go through a

times permeance of the magnetic circuit times $4\pi/10$), and, as every flux line links with every turn, the number of flux linkages is $1.257IN^2P$, which when substituted in Eq. 4 gives Eq. (8). This derivation is approximate in that it neglects the small amount of magnetic flux that does not stay confined to the magnetic material and so gives results that are slightly low. The additional inductance contributed by these flux lines can be calculated with fair accuracy by estimating the flux linkages contributed by this extra flux. Calculations of this type are described at length in many books dealing with magnetic-circuit calculations, as, for example, Karapetoff, *op. cit.*

displaced minor hysteresis loop that is superimposed upon the usual hysteresis curve. Examples of such displaced hysteresis loops are shown at 1, 2, and 3 in Fig. 8. In this figure one tip of each minor loop is shown resting on the normal loop. Actually, after the minor loop has been traversed by a number of cycles of magnetization, the minor loop gradually shifts its position and ceases to touch the main loop. The permeance of the core to the alternating current that is superimposed upon the direct current, and hence the effective coil inductance offered to the alternating current, is proportional to the slope of the line (shown dotted in Fig. 8) joining the two tips of the displaced hysteresis loop. The permeability shown by a magnetic material to alternating currents superimposed upon direct currents is called the *incremental permeability* (*i.e.*, the permeability to a small increment of alternating magnetomotive force) and has been the subject of considerable study.[1] The most important characteristics of incremental permeability (and hence of the inductance to superimposed alternating currents) are: (1) for a given alternating current the incremental permeability (and hence the inductance) to the superimposed alternating current will be less the greater the direct current upon which the alternating current is superimposed; and (2) with a given direct current the incremental permeability, and hence the inductance to the alternating current, will increase as the superimposed alternating current becomes larger. These characteristics hold until the flux density becomes high, and are clearly brought out by Fig. 9.

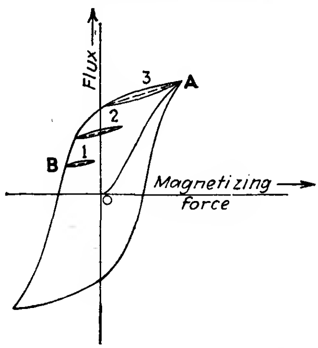


FIG. 8.—Typical hysteresis loop showing displaced minor hysteresis loops 1, 2, and 3, obtained when a small alternating current is superimposed upon a direct-current magnetization. The effective permeability to the superimposed alternating current is proportional to the slope of the line joining the tips of the displaced hysteresis loop.

When an alternating current of one frequency is superimposed upon a current of another frequency instead of upon direct current, the situation becomes very complicated. The general result is that the effective inductance to each frequency is altered by the presence of the other current; furthermore the inductance offered to each frequency component varies cyclically at the frequency of the other component.

When it is necessary to have an inductance that does not change appreciably with the current flowing through the windings, or when it is desired to minimize the various effects that result when the magnetic flux is not proportional to the coil current, or when the effects of direct-current saturation on incremental inductance are to be minimized, an air

[1] See Thomas Spooner, Permeability, *Trans. A.I.E.E.*, vol. 42, p. 340, 1923.

gap is left in the magnetic circuit. The total reluctance of the magnetic circuit of the coil is then the sum of the reluctances of the air gap and of the magnetic part of the magnetic circuit; by making the former much the largest (from five to twenty-five times that of the latter), variations in the reluctance (or permeability) of the core material have only a small effect on the total reluctance of the magnetic circuit. The presence of an air gap reduces the detrimental effects of the magnetic material but does not eliminate them, and the length of gap to use therefore depends upon the particular situation involved.[1]

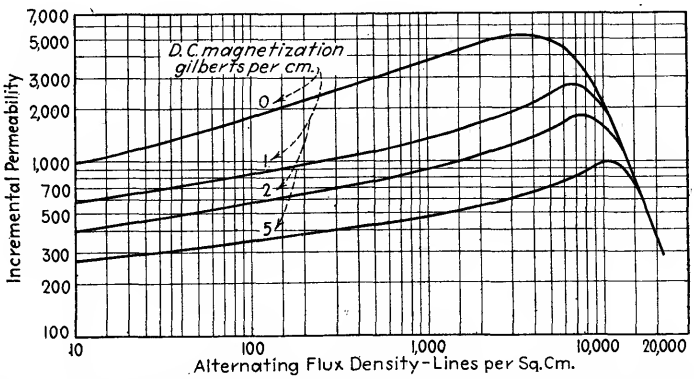


FIG. 9.—Curve giving incremental permeability of typical silicon steel as a function of alternating-current magnetization for several values of superimposed direct current, showing how the inductance decreases with increased direct current and reduced alternating-current magnetization.

The length of air gap usually required is only a small fraction of the total length of the magnetic circuit because the permeability of the magnetic material is hundreds of times that of air. Thus, although an air gap in the magnetic material reduces the coil inductance to a much smaller value than would be obtained with no gap, enough magnetic material is present in the coil to increase greatly the permeance of the magnetic circuit, and hence the inductance, over the value obtainable with no magnetic material whatsoever in the core. The necessary air gap can be supplied in a number of ways. Under some conditions the air gaps at the joints in the laminations are sufficient. In other cases a non-magnetic spacer is placed at a butt joint to increase the gap, while sometimes the cores are made up of insulated particles of powdered magnetic material pressed together to form a core that has minute air gaps distributed throughout its structure.

[1] The design of reactors carrying a superimposed direct current can be carried out by the method of C. R. Hanna, "Design of Reactances and Transformers Which Carry Direct Current," *Trans. A. I. E. E.*, vol. 46, p. 155, 1927.

Magnetic Materials.—Cores for such communication applications as audio-frequency transformers, power transformers, and filter reactors are most commonly of silicon-steel laminations. Special grades of such laminations have been developed which have lower losses and higher incremental permeability than ordinary silicon steel. In addition, a number of alloys having unusual magnetic properties are coming into increasing use in communication work. Chief among these are a series of nickel-iron alloys characterized by extremely high permeability and low hysteresis loss at low flux densities. Different compositions go under such names as permalloy, mu-metal, hypernik, nicoloi (or electric metal). A related nickel-iron-cobalt alloy called *perminvar,* which has a moderately high permeability at low flux densities, is remarkable for a permeability that is practically constant for magnetomotive forces up to about two ampere turns per centimeter, coupled with an extremely low hysteresis loss. A wide variety of other alloys are available for other special uses, such as permanent magnets, high permeability at high flux densities, etc.[1]

7. Mutual Inductance and Coefficient of Coupling.—When two inductance coils are so placed in relation to each other that flux lines produced by current in one of the coils link with the turns of the other coil as shown in Fig. 10*a*, the two inductances are said to be inductively coupled. The effects that this coupling produces can be expressed in terms of a property called the mutual inductance, which is defined by the relation:

$$\text{Mutual inductance } M \text{ in henries} = \frac{\begin{pmatrix}\text{flux linkages in second coil}\\ \text{produced by current in first coil}\end{pmatrix}}{\text{current in first coil}} 10^{-8} \qquad (9a)$$

$$= \frac{\begin{pmatrix}\text{flux linkages in first coil}\\ \text{produced by current in second coil}\end{pmatrix}}{\text{current in second coil}} 10^{-8} \qquad (9b)$$

Formulas (9*a*) and (9*b*) are equivalent and give the same value of mutual inductance. The flux linkages produced in the coil that has no current in it are counted just as though there were a current in this coil, so that the number of times a flux line would encircle an imaginary coil current is the number of linkages contributed by this particular line. In adding up the flux linkages it is important to note that different flux lines may

[1] For further information on special magnetic alloys see: G. W. Elmen, Magnetic Alloys of Iron, Nickel, and Cobalt, *Elec. Eng.*, vol. 54, p. 1292, December 1935; H. D. Arnold and G. W. Elmen, Permalloy, a New Magnetic Material of Very High Permeability, *Bell System Tech. Jour.*, vol. 2, p. 101, July, 1923; G. W. Elmen, Magnetic Properties of Perminvar, *Bell System Tech. Jour.*, vol. 8, p. 21 January, 1929; I. C. Pettit, Magnetic Materials, *Bell Lab. Record*, vol. 13, p. 39, October, 1932; "Standard Handbook for Electrical Engineers," Sec. 4, McGraw-Hill Book Company, Inc.

link with the same coil in opposite directions, in which case the total number of linkages is the difference between the sums of positive and negative linkages. The mutual inductance may therefore be positive or negative depending upon the direction of the linkages.

The problem of calculating mutual inductance is similar in all respects to the problem of computing inductance, and formulas have been worked out by which the mutual inductance can be calculated with good accuracy in all the ordinary types of configurations. Some of the more important cases are treated in Appendix A.

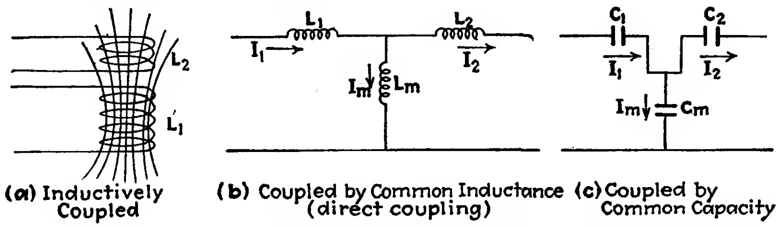


FIG. 10.—Several simple methods of coupling two circuits.

Coefficient of Coupling.—The maximum value of mutual inductance that can be obtained between two coils having inductances L_1 and L_2 is $\sqrt{L_1L_2}$. The ratio of mutual inductance actually present to this maximum possible value is called the coefficient of coupling, which can therefore be expressed by the relation:

$$\text{Coefficient of coupling} = k = \frac{M}{\sqrt{L_1L_2}} \tag{10}$$

The coefficient of coupling is a convenient constant because it expresses the extent to which the two inductances are coupled, independently of the size of the inductances concerned. In air-cored coils, as used in radio, a coupling coefficient of 0.5 is considered high and is said to represent "close" coupling, while coefficients of only a few hundredths represent "loose" coupling.

When two coils of inductance L_1 and L_2, between which a mutual inductance M exists, are connected in series, the equivalent inductance of the combination is $L_1 + L_2 \pm 2M$. The term $2M$ takes into account the flux linkages in each coil due to the current in the other coil. These mutual linkages may add to or subtract from the self-linkages, depending upon the relative direction in which the current passes through the two coils. When all linkages are in the same direction, the total inductance of the series combination exceeds by $2M$ the sum of the individual inductances of the two coils; while if the mutual linkages are in opposition to the other linkages, the inductance of the combination is less than the sum of the two coil inductances by $2M$. This property can be taken

advantage of to measure the mutual inductance between two coils. The procedure is to connect the two coils in series and measure the equivalent inductance of the combination. The connections to *one* of the coils are then interchanged and the equivalent inductance is measured again. The difference between the two measured inductances is then $4M$.

Any two circuits so arranged that energy can be transferred from one to the other are said to be coupled, even though this transfer of energy takes place by some means such as a condenser, resistance, or inductance common to the two circuits rather than by the aid of a mutual inductance. Examples of various methods of coupling are shown in Fig. 10. *Any two circuits that are coupled by a common impedance have a coefficient of coupling that is equal to the ratio of the common impedance to the square root of the product of the total impedances of the same kind as the coupling impedance that are present in the two circuits.* Thus with case *b* in Fig. 10, where the coupling is furnished by the common inductance L_m, the total inductances of the two circuits are $L_1 + L_m$ and $L_2 + L_m$, and the coefficient of coupling is given by the equation

$$\text{Coefficient of coupling } k \text{ for Fig. 10}b = \frac{L_m}{\sqrt{(L_1 + L_m)(L_2 + L_m)}} \tag{11}$$

In Fig. 10*c* the coupling element is a common capacity C_m, and the coefficient of coupling is[1]

$$\text{Coefficient of coupling for Fig. 10}c = \frac{\sqrt{C_1 C_2}}{\sqrt{(C_m + C_1)(C_m + C_2)}} \tag{12}$$

8. Condensers and Dielectrics.—Electrostatic capacity exists wherever an insulator (*i.e.*, dielectric) separates two conductors between which a difference of potential can exist. Applying a voltage between these conductors causes an electric charge to flow into them, with the resulting production in the dielectric of an electrostatic stress that repre-

[1] Equation (12) is readily derived by following the rule that the coefficient of coupling is equal to the ratio of common impedance to the square root of the product of primary and secondary impedance of the same kind as the coupling element. Thus, in Fig. 10*c*, the capacity of the primary is $\frac{C_1 C_m}{C_1 + C_m}$, and that of the secondary is $\frac{C_2 C_m}{C_2 + C_m}$. The coupling reactance is $1/\omega C_m$, while the primary and secondary reactances are $\frac{C_1 + C_m}{\omega C_1 C_m}$ and $\frac{C_2 + C_m}{\omega C_2 C_m}$, respectively. The coefficient of coupling is then

$$k = \frac{1/(\omega C_m)}{\sqrt{\frac{C_1 + C_m}{\omega C_1 C_m} \cdot \frac{C_2 + C_m}{\omega C_2 C_m}}}$$

which reduces to Eq. (12).

sents stored electrical energy.[1] The combination of the conducting plates separated by the insulating dielectric is called a condenser. The amount of energy stored in this way in the electrostatic field of the condenser depends upon the voltage that is applied to the electrodes and upon the electrostatic capacity of the combination. The capacity increases directly with the area and inversely with the thickness of the dielectric, and depends to a considerable extent upon the kind of dielectric being used. Electrostatic capacity is measured in farads, which is a capacity of such a size that, when 1 volt is applied across a capacity of 1 farad, a charge of 1 coulomb (equivalent to 1 amp. of current flowing for 1 sec.) will be stored. The farad is a very large unit and is frequently subdivided into the microfarad (abbreviated μf).

The principal properties of condensers are summarized in the following equations, in which C is in farads and E in volts.[2] The energy stored in a capacity C when charged to a voltage E is

$$\text{Stored energy in joules} = \tfrac{1}{2}CE^2 \tag{13}$$

The charge stored in a capacity C when charged to a voltage E is

$$\text{Stored charge in coulombs} = CE \tag{14}$$

The current that flows into or out of the condenser at any instant is the rate of change of charge, and so is given by the relation

$$\text{Current flowing into condenser} = \frac{dQ}{dt} = C\frac{dE}{dt} \tag{15}$$

When the voltage is sinusoidal, $C\dfrac{dE}{dt}$ reduces to $j\omega CE$. In the case where the condenser electrodes are large plates having a constant spacing, the capacity is given by the equation

$$\text{Capacity in } \mu\mu\text{f} = 0.08842K\frac{A}{d} \tag{16}$$

where

A = area of active dielectric in square centimeters
d = spacing between plates in centimeters
K = constant, called the dielectric constant, which depends upon the dielectric material and is substantially independent of frequency

Values of K for common dielectrics are given in Table II.

[1] For many years it was thought that the charge resided in the dielectric instead of the conductors, but it has been recently discovered that in circumstances where the charge appeared to exist in the dielectric it was really on the surface of the dielectric in a conducting moisture film that acted as an extension of the electrodes.

[2] The derivation of these equations is to be found in most textbooks of physics and electrical engineering and will therefore not be given here.

Losses in Condensers.—A perfect condenser when discharged gives up all the electrical energy stored in it, but this ideal is never completely realized since all actual condensers return less energy than was expended in charging them. The energy not delivered by the discharge represents losses that cause the condenser to have a power factor that is not zero. Practically all the electrical losses of the usual condenser are due to dielectric hysteresis, which is the result of molecular friction in the dielectric and is similar in character to magnetic hysteresis. The energy lost in the condenser as a result of dielectric hysteresis appears in the form of heat generated within the dielectric.

TABLE II.—CHARACTERISTICS OF DIELECTRICS AT RADIO FREQUENCIES WITH NORMAL ROOM TEMPERATURE[1]

Material	Dielectric constant	Power factor
Air	1.00	0.000
Mica (electrical)	5 to 9	0.0001 to 0.0007
Hard rubber	3 to 5	0.006 to 0.014
Glass (electrical)	4.90 to 7.00	0.004 to 0.016
Bakelite derivatives	4.5 to 7.5	0.02 to 0.09
Celluloid	4.10	0.042
Fiber (dry)	4 to 6	0.02 to 0.09
Wood (without special preparation)		
Oak	3.3	0.039
Maple	4.4	0.033
Birch	5.2	0.065
Mycalex	8	0.002
Isolantite	6.1	0.0018
Porcelain (wet process)	6.5 to 7.0	0.006 to 0.008

[1] These data were compiled from various sources and represent typical values that can be expected.

The power factor of a condenser is determined by the type of dielectric used and is practically independent of the condenser capacity, the applied voltage, the voltage rating, or the frequency. This constancy of the power factor under such widely different conditions makes the power factor of a dielectric of fundamental importance where the losses in the condenser are to be considered. While the power factor of a condenser is determined by the dielectric used, it is also affected by the conditions under which the dielectric operates. Thus the power factor always becomes higher as the temperature is raised. Moisture absorbed in the dielectric also has an extremely unfavorable effect on the power factor, and with dielectrics such as paper and fiber, which readily absorb moisture, the power factor may become extremely poor if the humidity is appreciable.

Table II gives information relative to the power factor of various dielectrics commonly used in radio work. In Table III the effect that changes in frequency and temperature have upon the power factor and dielectric constant is given for several dielectrics. The reason for the constancy of the power factor with changes in frequency is that with the commonly used dielectrics the dielectric loss per cycle is almost independent of the number of cycles per second, so that, irrespective of the frequency, a constant proportion of the energy that is supplied to the condenser disappears as dielectric hysteresis.

TABLE III.—EFFECT OF FREQUENCY AND TEMPERATURE ON DIELECTRIC CHARACTERISTICS[1]

Material	Temperature, degrees centigrade	Frequency, kilocycles	Dielectric constant	Power factor
Bakelite derivative	21	500	5.6	0.054
	71	500	6.9	0.11
	120	500	10.4	0.38
Hard rubber	21	500	3.0	0.009
	71	500	3.1	0.021
	120	500	3.2	0.065
Bakelite derivative	21	295	5.9	0.051
	21	500	5.8	0.051
	21	670	5.7	0.051
	21	1040	5.6	0.058
Hard rubber	21	210	3.0	0.009
	21	440	3.0	0.009
	21	710	3.0	0.009
	21	1126	3.0	0.010

[1] E. T. Hoch, Power Losses in Insulating Materials, *Bell System Tech. Jour.*, vol. 1, p. 110, November 1922.

The losses of a dielectric are sometimes expressed in terms of the angle by which the current flowing into the condenser fails to be 90° out of phase with the voltage. This angle is called the *phase angle* of the dielectric, and its value in radians is equal to the power factor. Thus a power factor of 0.01 represents a phase angle of 0.01 radian or 0.573°.

Equivalent Series and Shunt Resistance.—While a comparison of the losses of different dielectrics can be most satisfactorily expressed in terms of power factor or phase angle, the effect that the losses of a particular condenser have on the associated electrical circuits can be taken into account by replacing the actual condenser with a perfect condenser of the same capacity and having a resistance in series as shown in Fig. 11*b* or a resistance in parallel as in Fig. 11*c*. The value of the series or

shunt resistance is so selected that the power factor of the perfect condenser associated with the resistance is the same as the power factor of the actual condenser. The value of the series resistance can be computed in terms of the power factor, condenser capacity C, and frequency f in the usual way and is given to a high degree of accuracy by the equation

$$\text{Series resistance} = R_1 = \frac{\text{power factor}}{2\pi f C} \tag{17}$$

In the same way the shunt resistance that can be used to represent the actual losses of the condenser is related to the power factor, capacity,

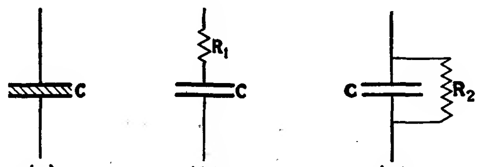


FIG. 11.—Representation of imperfect condenser by a perfect condenser of same capacity with series resistance and by a perfect condenser with shunt resistance.

and frequency to a high degree of accuracy by the equation

$$\text{Shunt resistance} = R_2 = \frac{1}{(2\pi f C)(\text{power factor})} \tag{18}$$

The relationship between the shunt resistance and series resistance for a given condenser at a given frequency can be obtained by combining Eqs. (17) and (18) to eliminate the power factor, which yields the result

$$\left.\begin{aligned} R_1 &= \frac{1}{R_2(2\pi f C)^2} \\ R_2 &= \frac{1}{R_1(2\pi f C)^2} \end{aligned}\right\} \tag{19}$$

Both the equivalent series and shunt resistances of a condenser having a constant power factor are independent of the voltage applied to the condenser but vary inversely with the condenser capacity and the frequency.

In addition to dielectric hysteresis, it is possible for other losses to exist in a condenser. Among these are the losses caused by resistance of the condenser plates, the loss from leakage currents passing through the dielectric, and losses arising from corona discharges. The ohmic loss is small but is not always negligible at very high frequencies. The effects of leakage are wholly negligible at ordinary frequencies and become of importance only when a condenser is used to block off a direct current, in which case even a small leakage may be of importance. The

leakage of a condenser is greatly increased by moisture in the dielectric and by raising the temperature of the dielectric. The leakage is ordinarily expressed in terms of the equivalent resistance between the condenser terminals and is inversely proportional to the capacity of the condenser. Corona occurs only in high-voltage condensers and is to be avoided in all circumstances as it represents a large power loss and in addition starts chemical actions that cause rapid deterioration of all solid dielectrics in the vicinity.

Condensers with solid dielectrics are ordinarily freed of moisture and then impregnated with some type of insulating compound during manufacture. This has the effect of making the condenser moistureproof and thus insures that the power factor and leakage will be maintained at a low value irrespective of weather conditions. In high-voltage condensers such impregnation, when done by a vacuum process, also has the effect of eliminating air spaces between the dielectric and the condenser plates, which removes the possibility of corona and therefore greatly increases the voltage that may be safely applied to the condenser.

Condensers for Radio Purposes.—Paper, mica, and air are the dielectrics most frequently employed in condensers used in radio work. Paper dielectric is inexpensive and gives a large capacity in a small volume but has the disadvantage of fairly high electrical losses. Paper condensers are usually constructed by insulating two strips of tin or copper foil with enough layers of paper strips to give the necessary voltage strength and then rolling these into a compact bundle which is made moistureproof by sealing with insulating compound. The dielectric losses and the insulation strengths of different grades of paper vary widely, and unless paper especially made for condenser use is employed, the leakage and dielectric hysteresis will be high and the voltage rating low.

Mica is widely used as a dielectric in high-grade condensers because of its high voltage strength and its superiority over practically all other solid dielectrics in the matter of losses. Mica is expensive, however, and is not used where paper is satisfactory (*i.e.*, when the losses are unimportant), except when the condenser capacity is so small that the cost of the dielectric is insignificant. Mica condensers are constructed by piling alternate layers of metallic and mica sheets upon one another and then connecting alternate layers of conductors together to form the two electrodes of the condenser. This type of construction is necessary because mica is damaged when rolled. The quality of the mica used is very important, and only the best grades are satisfactory for electrical purposes.

Air is a perfect dielectric, having no dielectric hysteresis loss whatsoever. The principal losses that condensers with air dielectric have when operating below the corona voltage are those introduced by the dielectric

material used in mounting the plates. Air condensers are extremely bulky in proportion to capacity because of the low dielectric constant of air and because a relatively large clearance between plates is necessary to insure separation in view of mechanical irregularities.

Although paper, mica, and air are the dielectrics generally used in condensers, the dielectric losses in other materials such as bakelite, hard rubber, fiber, porcelain, and various molded insulating compounds are important because these materials are often used in one form or another as insulators. The solid dielectric used in the air condenser to separate the two sets of plates is an excellent example of such use. Many other cases may be cited, such as tube bases, tube sockets, coil forms, panels, and so on. Although the dielectric losses that are present in such cases are generally small because of the small quantity of dielectric involved, they are sometimes very important, as is the case with the low-loss variable condenser used in tuning circuits to resonance, the requirements of which are discussed below.

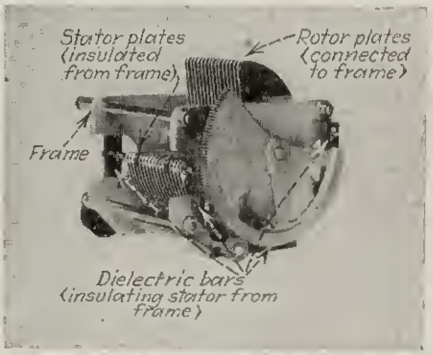


Fig. 12.—A typical laboratory type of air-dielectric variable condenser showing details of construction. In this condenser the rotor plates are connected to the end plates.

Variable Condensers with Air Dielectric.—The chief use of air as a dielectric is in condensers used to adjust the resonant frequency of tuned circuits, where a small condenser having low losses at radio frequencies and capacity that is continuously variable is required. Air-dielectric condensers for this purpose are usually constructed as shown in Fig. 12, with a series of fixed plates, spaced by washers, as one electrode, and a series of similarly spaced rotating plates as the other electrode. The rotating plates mesh with the fixed plates without touching them, thus causing the condenser to have a capacity determined by the angle of rotation. The way in which the capacity of the variable condenser varies with the angle of rotation can be controlled either by cutting the stationary or the rotating plates to a special shape or by use of an air gap between the two sets of plates which varies with the angle of rotation. In order to insure that the condenser will have low losses, the dielectric used to insulate the fixed and moving plates from each other should be arranged in a way that will keep the voltage gradient in it low and should have the lowest possible power factor. Hard rubber is widely used for this purpose and is superior to bakelite, which has a much greater power factor. The details of construction of a typical condenser can be seen

in Fig. 12. In this condenser the metal end plates are connected to the moving plates by a hair-spring and act as a shield for the fixed plates, from which they are insulated by the strips of dielectric. The usual variable air-dielectric condenser has a capacity of 0.0005 μf or less. Larger sizes are rather bulky, and air condensers with capacities in excess of 0.002 μf are rarely made.

Condensers in which the fixed and rotating plates are both semicircular have a capacity that is roughly proportional to the angle of rotation and

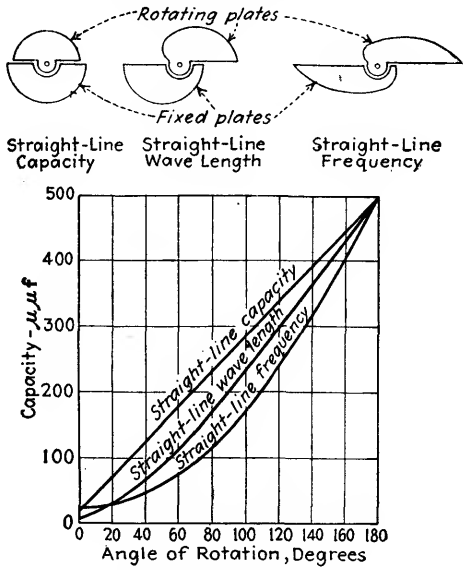


FIG. 13.—Characteristics of straight-line capacity, straight-line-wave-length, and straight-line-frequency condensers, showing capacity as a function of angle of rotation and the approximate shape of plates for each type.

are therefore called straight-line-capacity condensers. For some purposes it is desirable to have a condenser in which the capacity is proportional to the square of the angle of rotation. Such a condenser, when used to resonate with a fixed inductance, causes the wave length at resonance to be proportional to the angle of rotation, and so is called a straight-line-wave-length condenser. Still other types of condensers are so constructed that the resonant frequency is proportional to the angle of rotation, thus giving a straight-line-frequency type of condenser. The shape of the plates required in the straight-line-wave-length and straight-line-frequency condensers is affected by the distributed capacity of the coil that is being tuned by the condenser, so that in order to get absolute straight-line-wave-length or straight-line-frequency characteristics it is necessary to design the condenser to match a particular coil. The approximate shapes of rotor plates that will give these effects are given in Fig. 13,

which also gives the variation of capacity with angle of rotation for these types of condensers.[1]

In a single-dial-control radio receiver it is necessary to tune several circuits simultaneously. This is accomplished by using a variable condenser having several sets of rotating plates all mounted on a common shaft, with a set of insulated stator plates to go with each rotor section. When the various circuits involved are all to be resonant at the same frequency, the different sections of the condenser are made as nearly alike as possible, and then any residual inequalities are eliminated by bending the plates and by using small auxiliary "trimmer" condensers which are associated with each section and which can be adjusted by a screw driver. When the different sections are to tune circuits to different but related frequencies, as in the case of the superheterodyne receiver, one can either make all sections alike and then design the associated circuits to give the required frequency difference, or can make the plates of the different sections of different shape. This is discussed further in Sec. 110.

Electrolytic Condensers.—The electrolytic condenser makes use of the fact that aluminum (and certain other materials), when placed in a suitable solution and made the positive electrode, will form a thin insulating surface film which will withstand a considerable voltage and which at the same time will have a high electrostatic capacity per unit area of film because of the extreme thinness of the film. This film is the result of electrochemical action and is formed by applying positive voltage to the aluminum electrode. The thickness of the film and hence the capacity obtained depend largely upon the voltage used in this forming process, with the lower voltages giving the greatest capacity. The maximum voltage that the film can be made to withstand depends somewhat upon the electrolyte used, and can be as great as 600 volts. The composition of the electrolyte varies with different types of electrolytic condensers, and may be a liquid, a highly viscous fluid, or a paste, giving so-called wet, semidry, and dry condensers, respectively. The negative electrode is commonly supplied by the metal container, and is sometimes, although not always, of aluminum.

Electrolytic condensers are characterized by a high capacity per unit volume. They also have much higher leakage current than do other condensers, and their power factor is high and dependent upon frequency. The capacity, power factor, leakage, etc., are not particularly stable, changing with age, applied voltage, and temperature. When excessive voltage is applied, the insulating film breaks down, but tends to reform when the voltage is lowered as a result of electrochemical action. If

[1] Formulas for calculating the shape of plates in straight-line-frequency and wavelength condensers are to be found in O. C. Roos, Simplified S.L.F. and S.L.W. Design, *Proc. I.R.E.*, vol. 14, p. 773, December, 1926.

voltage is applied to the condenser in the reverse polarity, the condenser acts as a low resistance and passes current freely.

Electrolytic condensers are widely used in radio work for filter and by-pass purposes. In such applications they are subjected to a steady direct-current voltage, and are then called upon to short-circuit any superimposed alternating potential. Under such conditions the exact capacity and the losses are unimportant. In proportion to capacity and voltage rating, electrolytic condensers are the least expensive condensers available, but they have the disadvantage of deteriorating with time as a result of destructive chemical action.

Voltage Rating of Condensers.—The voltage that can be safely applied to condensers having solid dielectrics such as paper and mica depends upon the insulating strength of the dielectric used and upon the electrical losses in the dielectric. If the applied voltages exceed the dielectric strength, the dielectric will spark through and the condenser will be destroyed. However, if the losses in the condenser are sufficient to cause a moderate amount of heating, the allowable voltage will be something less than the dielectric strength of the material because of the fact that all dielectrics deteriorate rapidly when heated. As the condenser losses are proportional to frequency, the voltage rating of a particular condenser will be highest on direct-current potentials, somewhat less at low frequencies, and increasingly lower as the frequency is increased, until at high radio frequencies the allowable voltage is inversely proportional to the square root of the frequency and is only a small fraction of the insulation strength to direct current. It is therefore very important that condensers which must withstand high radio-frequency voltages have low losses. Otherwise the heat that is generated in the dielectric will raise the temperature to a point where deterioration and eventual destruction will take place.[1]

[1] The importance of losses in determining the voltage rating is illustrated by the following ratings of a particular 0.001 μf mica condenser as given by I. G. Maloff, Mica Condensers in High Frequency Circuits, *Proc. I.R.E.*, vol. 20, p. 647, April, 1932:

Frequency, kc	Rated effective voltage, volts
Direct current	10,000
1	10,000
100	3,000
300	3,000
1,000	1,780
3,000	605
10,000	178

Mica, because of its very low losses and high insulation strength, is the dielectric universally used for high-voltage service at radio frequencies. It is very important that high-voltage condensers with solid dielectrics be thoroughly impregnated to eliminate corona and the losses resulting from moisture. Transformer oil and some other types of oil can be used as a dielectric for high-voltage service but these have the disadvantages of picking up moisture from the atmosphere and of possessing an insulating strength in bulk that is relatively low compared with the value in small quantities. Variable condensers with air dielectric are occasionally used in high-voltage radio-frequency service. Such condensers are extremely bulky in proportion to their capacity because the spacing between plates must be very large if the voltage rating is to be reasonably high. The result is that high-voltage air condensers are used only at high radio frequencies where the capacity is very small. The voltage that may be applied to an air-dielectric condenser is proportional to the air pressure; so enclosing the condenser in a container filled with air at a pressure of from 10 to 20 atmospheres greatly increases the potential that may be safely applied.

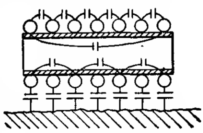


FIG. 14.—Some of the coil capacities that contribute to the distributed capacity of a single-layer coil.

Inductance of Condensers.—All condensers have a certain amount of inductance in their circuit because of the fact that the current flowing through the condenser produces magnetic flux. The major part of this inductance occurs in the lead wires of the condenser, but a certain amount results from magnetic flux set up by the current in the condenser plates. The inductance of condensers is very small even under the most unfavorable conditions and is therefore important only when the frequency is extremely high. Condenser inductance can be minimized by properly arranging the lead wires and the connections to the condenser plates.

Distributed Capacity of Coils.—Every electric circuit and every piece of electrical equipment has capacity associated with it because there are always dielectrics separating conductors between which voltage exists. These capacities are very often quite small, but at very high frequencies even a small condenser has a low reactance and so becomes important. The stray capacity of an inductance coil is an important example of this.

In an air-core inductance coil there are small capacities between adjacent turns, capacities between turns that are not adjacent, and capacities between terminal leads. In addition there can be capacities to ground from each turn. Some of the different capacities that may exist in a typical coil are shown in Fig. 14. It is to be noted that every turn has a capacity to every other turn and also a capacity to ground. Each of these capacities stores a quantity of electrostatic energy that is determined by the capacity and the fraction of the total coil voltage that

appears across the turns involved. The total effect which the numerous small coil capacities have can be represented to a high degree of accuracy by assuming that these many capacities can be replaced by a single condenser of appropriate size shunted across the coil terminals. This equivalent capacity is called either the distributed capacity or the self-capacity of the coil.

In multilayer coils the distributed capacity tends to be high because the layer arrangement of the winding causes turns from different parts of the winding to be located near each other. Thus in the two-layer winding shown at Fig. 15*a*, in which the turns are numbered in order, the first and last turns are adjacent, which causes the voltage between these turns to be large and hence greatly increases the effect of the shunting capacity. An improvement over the simple layer winding of Fig. 15*a* is obtained by using the type of winding shown in Fig. 15*b*, known as a *bank winding*, in which adjacent turns represent parts of the coil that are close together, while the ends of the windings are fairly far apart. Another method of keeping down the self-capacity of a multilayer coil is to use many layers with only a few turns per layer, as in Fig. 15*c*, which has the effect of separating the ends of the coil. Still another device for reducing self-capacity is shown in Fig. 15*d*, in which the layers are spaced to reduce the capacity from turns in one layer to turns in adjacent layers. Many ingenious methods have been devised for incorporating these principles into types of construction having desirable mechanical rigidity, and several types of multilayer coils with low capacity are produced commercially.

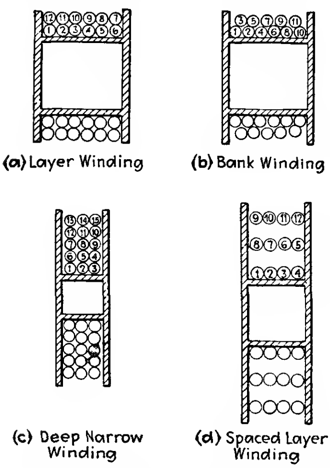


FIG. 15.—Several types of multilayer windings.

The various coil capacities that contribute toward the distributed capacity include as part of their dielectric the wire insulation and the form upon which the coil is wound. Since these materials have appreciable dielectric hysteresis, there is a loss associated with the distributed capacity of the coil which is referred to as the dielectric loss of the coil and which has the effect of increasing the effective coil resistance. In order to keep the dielectric loss small, the coil form should be of material having low dielectric losses and should be no thicker than mechanical considerations make advisable. It is also important that the winding be covered with a moistureproof binder such as collodion to prevent absorption of moisture by the cotton or silk insulation of the wire. The dielectric

losses are low and often negligible in well-constructed coils but may reach an undesirable magnitude if no attention is paid to them.

9. The Effective Resistance of Coils and Conductors at Radio Frequencies.—The effective resistance offered by conductors to radio frequencies is considerably more than the ohmic resistance measured with direct currents. This is because of an action known as skin effect, which causes the current to be concentrated in certain parts of the conductor and leaves the remainder to contribute little or nothing toward carrying the current. As a result of this effect *it is necessary to generalize the concept of resistance when dealing with radio frequencies by considering the resistance to be that quantity which when multiplied by the square of the current will give the energy dissipated in the circuit.*

Skin Effect in an Isolated Conductor.—A simple example of skin effect, and one which makes its nature clear, is furnished by an isolated round wire. When a current is flowing through such a conductor, the magnetic flux that results is in the form of concentric circles as shown in Fig. 16. It is to be noted that some of this flux exists within the conductor and therefore links with, *i.e.*, encircles, current near the center of the conductor while not linking with current flowing near the surface. The result is that the inductance of the central part of the conductor is greater than the inductance of the part of the conductor near the surface because of the greater number of flux linkages existing in the central region. At radio frequencies the reactance of this extra inductance is sufficiently great to affect seriously the flow of current, most of which flows along the surface of the conductor where the impedance is low rather than near the center where the impedance is high. The center part of the conductor therefore does not carry its share of the current and the effective resistance is increased, since in effect the useful cross section of the wire is very greatly reduced. The actual type of current distribution obtained in the case of a round wire is as shown in Fig. 16.

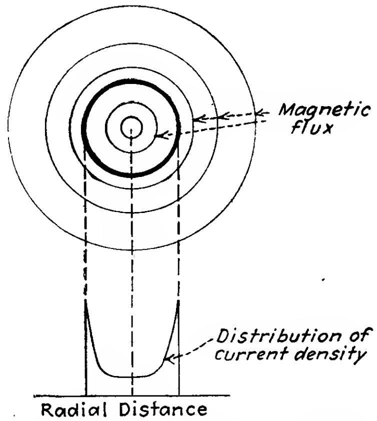


FIG. 16.—Isolated round conductor, showing flux paths and current distribution at radio frequencies. Note that the current density is highest for parts of the conductor encircled by the fewest flux lines.

The ratio that the effective alternating-current resistance bears to the direct-current resistance of a conductor increases with frequency, with conductivity of the conductor material, and with the size of conductor. This results from the fact that a higher frequency causes the extra inductance at the center of the conductor to have a higher reactance,

while a greater conductivity makes the reactance of the extra inductance of more importance in determining the distribution of current, and a greater cross section provides a larger central region. It is to be noted, however, that a larger conductor always has less radio-frequency resistance than a smaller one because, although the ratio of alternating-current to direct-current resistance is less favorable, this is more than made up by the greater amount of conductor cross section present. The actual value of the ratio of alternating-current to direct-current resistance for a solid round conductor can be calculated with the aid of tables and formulas found in every electrical handbook.

When skin effect is present, the current is redistributed over the conductor cross section in such a way as to make most of the current flow where it is

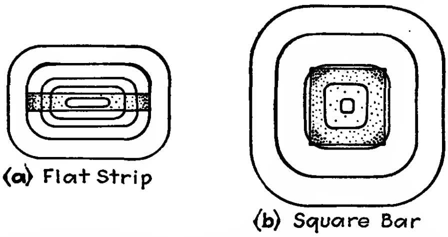


FIG. 17.—Isolated strip and bar conductors, showing approximate flux paths and current distribution. The current density is indicated by the density of shading and is seen to be greatest for those parts of the conductor encircled by the fewest flux lines.

encircled by the smallest number of flux lines. This general principle controls the distribution of current irrespective of the shape of the conductor involved. Thus with a flat-strip conductor, such as shown in Fig. 17, the current flows primarily along the edges, where it is surrounded by the smallest amount of flux, and the effective resistance will be high because most of the strip carries very little current. This illustration makes clear that it is not the amount of conductor surface that determines the resistance to alternating current but rather the way in which the conductor material is arranged. Another example is formed by the square-bar conductor of Fig. 17, in which the current is concentrated along the four corners as indicated in the figure because the flux linkages are least for this portion of the cross section.

Where it is important that an isolated conductor have very low resistance to radio-frequency currents it is preferable to use thin tubular conductors instead of round wire of the same cross section, since the central part of the wire carries very little current while all the material of a tube is effective. Another way of making more effective use of the conducting material is to form the conductor from a large number of small enameled wires connected in parallel at their ends but insulated from each other throughout the rest of their length, and thoroughly

interwoven. If the stranding is properly done, each conductor will, on the average, link with the same number of flux lines as every other conductor, and the current will divide evenly between the strands. If each strand is small, it will have relatively little skin effect over its cross section, with the result that all the material is effective in carrying the current and a radio-frequency resistance approximating the direct-current resistance results. A stranded cable of this type is called a *litz* (or *litzendraht*) conductor.

Skin Effect in Coils.—The same principle that governs the current distribution in an isolated conductor also determines the distribution of current in the conductors of a coil; that is, the current density is greatest in those parts of each coil conductor encircled by the smallest number of flux lines. The skin effect in coils is, however, much more complicated and much greater in magnitude than in isolated conductors because each turn of the coil produces flux that causes skin effect in adjacent turns. As a result the radio-frequency resistance of coils may be as much as several hundred times the resistance to direct currents. The approximate current distribution in the conductors of a typical radio coil is indicated by the shading in Fig. 6, which also shows the flux paths and so brings out the relation between current density and flux linkages.

The losses in a coil are most conveniently expressed in terms of the ratio of the coil reactance ωL to the effective coil resistance R. This ratio approximates the reciprocal of the coil power factor and is so important in the theory of resonant circuits that it is considered as a fundamental coil property and is usually referred to by the symbol Q; that is

$$Q = \frac{\text{coil reactance}}{\text{coil resistance}} = \frac{\omega L}{R} \tag{20}$$

The effective coil resistance R includes any dielectric loss that the coil may have.

It is convenient to express the characteristics of a coil in terms of the ratio of coil reactance to effective resistance because this ratio Q is approximately constant for the same coil over a wide range of frequencies as a result of the fact that the effective radio-frequency resistance of a coil is roughly proportional to frequency. Furthermore, the value of Q for equally well-designed coils intended for use at different frequencies is approximately the same, that is, a value of Q which is considered as denoting an efficient coil of a size suitable for use at 100 kc also represents the Q of an efficient coil for 1000-kc service. The values of Q actually obtained with coils used in receiving equipment range from fifty to several hundred, with values somewhat greater than these frequently encountered in transmitter inductances.

Factors Influencing the Q of a Coil.—The actual value of $\omega L/R$ depends primarily on the coil construction and is determined by complicated factors for which a complete mathematical solution has never been made. In spite of this there are a number of general principles that can be used as a guide in the design of coils. To begin with, if coils differing only in wire diameter are compared, it will be found that for each frequency there will be a particular size of wire that will give the lowest radio-frequency resistance and hence the highest Q. This best size of wire will be different for different frequencies and is often, although not necessarily, the largest wire that can be wound in the space available. As

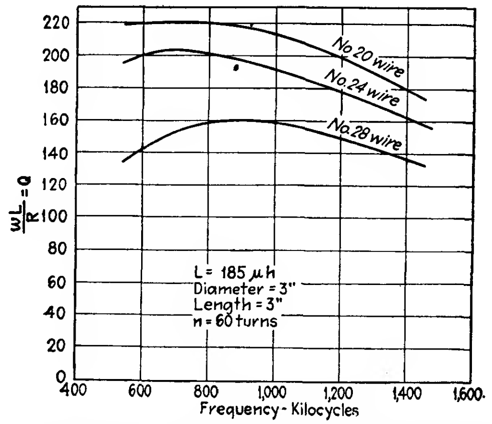


FIG. 18.—Curves showing values of Q as a function of frequency for three coils differing only in wire size. The largest size wire has over six times the cross section of the smallest, and yet the latter has a value of Q only about 35 per cent less.

long as the wire is not too small, changes in wire size have surprisingly little effect on coil alternating-current resistance because the increased skin effect of the larger conductors offsets in large measure the greater cross section. It is only when the frequency is so low or the wire so small as to give little skin effect that the radio-frequency resistance of the coil becomes markedly influenced by the direct-current resistance. These effects of wire size are graphically shown by the curves of Fig. 18, which give Q as a function of frequency for several coils differing only in wire size, and in which increasing the direct-current resistance over six times reduces the coil Q by about 35 per cent.

Increasing the size of a coil while maintaining the inductance, proportions, and direct-current resistance constant tends to reduce the effective radio-frequency resistance and hence give a larger Q. Furthermore the larger coil provides a winding space for larger wire than the smaller coil, which if utilized will still further increase the Q as the dimensions are increased. A large coil can utilize a large conductor to better

advantage than a small coil because the increased space over which the flux is distributed reduces the flux density in the vicinity of the conductors and hence reduces the skin effect to reasonable values even with large

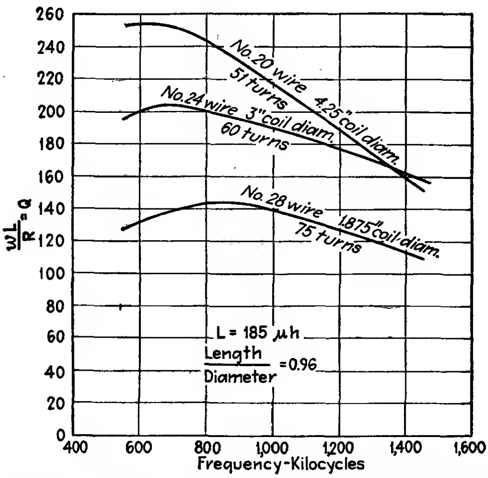


FIG. 19.—Curves showing values of Q as a function of frequency for three coils having the same inductance and same ratio of length to diameter, but differing in size. Increasing the size of the coil increases the Q at the frequency where the Q is highest and tends to lower the frequency at which the coil is most efficient.

wire. The effect that changes in size can have on the Q of a coil is demonstrated by the curves of Fig. 19.

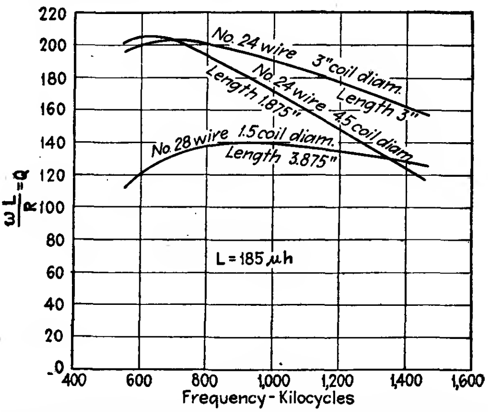


FIG. 20.—Curves giving values of Q as a function of frequency for three coils having the same inductance but differing in ratio of length to diameter. These show how there is a best shape.

The best shape for a coil having a given inductance is neither a very long coil with a small diameter nor a short coil with a large diameter but rather one of intermediate proportions. The curves of Fig. 20 are

typical examples of the differences that can be expected from different shapes.

The types of conductors most commonly used in coils are round wire, litz, tubing, and flat and edgewise-wound strip. Ordinary wire represents one of the best shapes and finds use at all frequencies, although litz wire is preferable at broadcast frequencies under some conditions and is very markedly superior to solid wire at lower frequencies (see Fig. 21). Tubing and strip conductors are used in transmitter coils that must carry large currents.

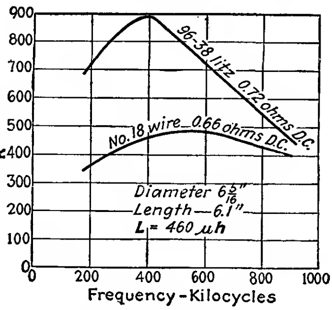


FIG. 21.—Curves giving values of Q as a function of frequency for two coils identical in every way except that one is wound with solid wire and the other with litz of approximately the same cross section. The litz coil has an exceptionally high Q.

The resistance of coils is normally obtained by measurements on the actual coils. Efforts to predict the resistance by calculation have been made by a number of investigators, with rather good success in some instances.[1] The difficulty is, however, that the theoretical formulas all involve some simplifying assumptions and also ignore dielectric losses, and so give results that are at best only approximate. Furthermore, the theoretical formulas are restricted to certain ranges of coil proportions, which reduces their usefulness.

Large quantities of data giving coil resistance in countless specific cases are to be found scattered throughout the literature, but no attempt has been made here to summarize the published results because they represent a huge mass of figures which have as yet defied organization in any systematic manner.[2] The important things to keep in mind are

[1] The most noteworthy work on this subject is that of Butterworth. See S. Butterworth, Effective Resistance of Induction Coils at Radio Frequency, *Exp. Wireless and Wireless Eng.*, vol. 3, pp. 203, 309, 417, and 483, April, May, July, and August, 1926. Explanatory discussions which are of value in applying Butterworth's results are given by B. B. Austin, Effective Resistance of Inductance Coils at Radio Frequency, *Wireless Eng. and Exp. Wireless*, vol. 11, p. 12, January, 1934; A. J. Palermo and F. W. Grover, Supplementary Note to Study of the High-frequency Resistance of Single Layer Coils, *Proc. I.R.E.*, vol. 19, p. 1278, July, 1931.

A review of other work that has been done, together with some original contributions, is given by A. J. Palermo and F. W. Grover, Study of the High Frequency Resistance of Single Layer Coils, *Proc. I.R.E.*, vol. 18, p. 2041, December, 1930.

[2] Useful information of this sort is to be found in the following publications: August Hund and H. B. De Groot, Radio-frequency Resistance and Inductance of Coils Used in Broadcast Reception, *Bur. Standards Tech. Paper* 298; E. L. Hall, Resistance of Conductors of Various Types and Sizes as Windings of Single-layer Coils at 150 to 6000 Kilocycles, *Bur. Standards Tech. Paper* 330; J. H. Morecroft, "Principles

the general principles discussed above and the fact that the effective resistance of a coil to alternating currents is determined primarily by the flux distribution in the conductors and has very little relation to the direct-current resistance of the wire.

10. Types of Coils Used in Radio Work. *Coils for Tuned Circuits of Radio Receivers.*—Coils used in resonant circuits of radio receivers must have low losses (*i.e.*, a high Q), small size, and reasonably small distributed capacity. The type of coil used to meet these requirements depends upon the frequency. At frequencies above about 1500-kc single-layer solenoids

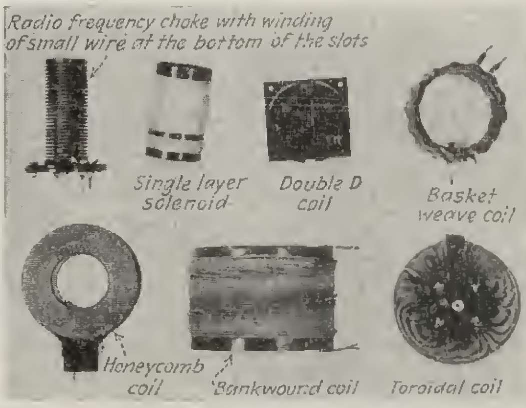


Fig. 22.—Representative coils of different types designed for use in radio receivers.

of solid wire are universally employed. The turns are nearly always spaced and are preferably wound upon a form such as a thin bakelite tube.

Coils for circuits resonant at broadcast frequencies are usually single-layer solenoids wound with solid wire. Bank-wound or "universal" (honeycomb) wound multilayer coils using litz wire are sometimes employed, and when properly designed will give a higher Q in the same volume than a single-layer coil.

Coils for frequencies below the broadcast range are nearly always bank-wound or universal multilayer coils, preferably using litz wire. Single-layer coils at these lower frequencies are seldom employed because of the large bulk necessary if the required inductance is to be obtained with reasonable size wire. Litz wire is much more satisfactory than solid wire at the lower frequencies and so is generally employed. Increasing use is also being made of iron dust in coils designed for frequencies below

of Radio Communication"; Coil Design for Short Wave Receivers, *Electronics*, vol. 7, p. 174, June, 1934; A. J. Palermo and F. W. Grover, A Study of the High Frequency Resistance of Single Layer Coils, *Proc. I.R.E.*, vol. 18, p. 2041, Dec. 1930; David Grimes and W. S. Barden, A Study of Litz Wire Coils, *Electronics*, vol. 6, pp. 303 and 342, November and December, 1933.

500 kc, and such dust-cored coils are often superior to air-cored coils occupying the same volume.

Small size is a very definite advantage in coils intended for radio receivers. Small coils make for compactness and also simplify greatly the shielding problem. Small size means increased losses, but it has been possible to design coils of very small size which have moderately high values of Q.

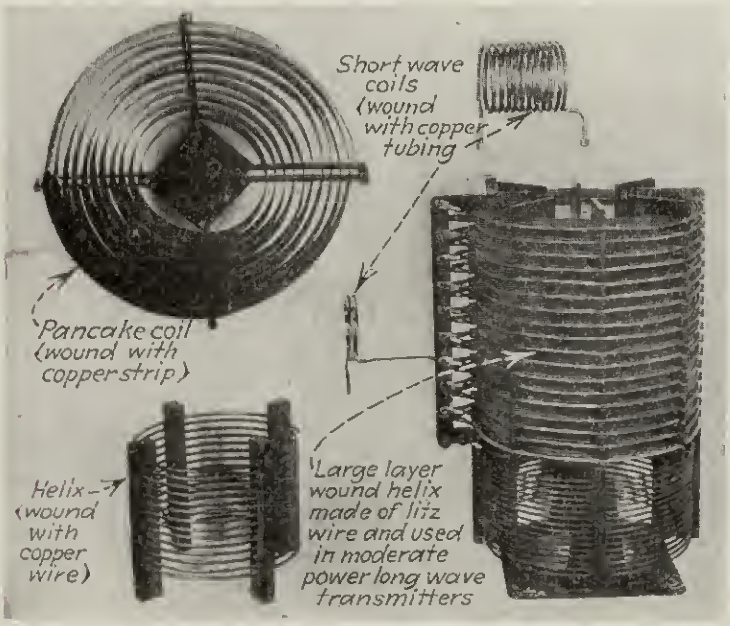


FIG. 23.—Representative coils of different types designed for use in radio transmitters.

A great many different types of coils have been used at one time or another in radio receivers. In some of these, such as the basket-weave coil, the aim has been to reduce the solid dielectric present. It has been found, however, that a thin solid form will not introduce appreciable losses and will greatly increase the coil rigidity. Other types of construction, such as the toroid and double D coil illustrated in Fig. 22, have as their object the elimination or reduction of the external magnetic field. Better results are obtained, however, by using a single-layer solenoid in a copper shielding can, since this arrangement gives both electrostatic and magnetic shielding and results in a higher Q for the same bulk.

Coils for Transmitters.—The method of coil construction and type of conductor employed in transmitters depend upon the frequency and power involved. Coils for short-wave transmitters are always made in the form of a single-layer solenoid. The conductor may be heavy wire, tubing, or a flat strip. Coils made of tubing are normally arranged to be

self-supporting as illustrated in Fig. 23, while those with wire and strip are either wound on a form, usually isolantite, or held in position with insulating strips as shown in the lower left-hand corner of Fig. 23. When the transmitter power is very high, the heat dissipated in the coil is considerable, and it is not uncommon to use coils constructed of tubing through which cooling water is passed.

Coils for broadcast transmitters are usually large single-layer solenoids wound with wire, copper tubing, or edgewise-wound copper strip. Pancake inductances of copper strip, large wire, or tubing are occasionally employed.

At the lower frequencies multilayer coils are usually employed in order to obtain the necessary inductance in a reasonable space, and litz wire, although much more expensive than large copper tubing, is very much better and so is rather generally preferred.

Some use is also being made of radically different types of inductances and tuned circuits to obtain very low losses. One method of approach is to use toroidal coils made with flat conductor so arranged that the flat side is parallel to the flux lines, thereby reducing the skin effect.[1] Another means of achieving low losses is to employ tuned circuits utilizing resonant transmission lines (see Sec. 15).

There are numerous design problems encountered in coils that are to handle large radio currents. In particular, the high voltages developed across such coils have a tendency to produce corona at the points where the electrostatic stress is highest. Since even the slightest trace of corona at radio frequencies results in high energy loss, great care must be used in arranging the details of the construction.[2]

Radio-frequency Choke Coils.—Certain types of radio circuits call for an inductance coil having a very high impedance to radio-frequency currents lying within a certain range of frequencies. The losses of such coils are quite unimportant, but the impedance, when taking into account the distributed capacity as well as the inductance, must be very high. Coils of this type are called radio-frequency choke coils and must have extremely low distributed capacity. The two methods commonly used in constructing such coils are shown at *a* and *b* of Fig. 24. In the first method the coil is very long in proportion to its diameter and has a large number of turns of small wire. The great length insures

[1] See F. A. Kolster, Generation and Utilization of Ultra-short Waves in Radio Communication, *Proc. I.R.E.*, vol. 22, p. 1335, December, 1934; F. E. Terman, Some Possibilities for Low Loss Coils, *Proc. I.R.E.*, vol. 23, p. 1069, September, 1935.

[2] An excellent discussion of the problems involved in the design of large inductances for use in high-power transmitting stations is to be found in W. W. Brown and J. E. Love, Design and Efficiencies of Large Air Core Inductances, *Proc. I.R.E.*, vol. 13, p. 755, December, 1925.

a small distributed capacity, and the many turns of small wire give a large inductance. In the second method the coil is wound in a series of spaced sections as shown in the figure. The over-all distributed capacity of the coil is kept small by making the multilayer winding in each section deep and narrow and by employing a number of sections in series. For best results the choke should behave as nearly as possible as an inductance shunted by a single capacity, in spite of the fact that the actual capacity is a combination of distributed and lumped elements. The desired mode of operation can be very closely approximated by a careful balance between the various capacities and the inductances of the various sections.[1] The resistance of radio-frequency coils is unimportant, and the size of wire is selected from the point of view of current-carrying capacity.

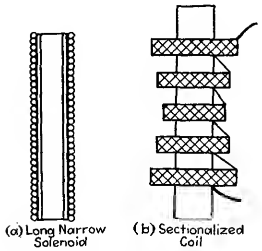


FIG. 24.—Radio-frequency choke coils.

Variable Inductances.—Inductances that are continuously variable can be constructed by connecting two coils in series and then varying the total inductance $L_1 + L_2 + 2M$ of the combination by changing the mutual inductance. Such variable inductances are called variometers and have many uses. When each of the two coils has an inductance of L and the maximum coefficient of coupling that can be obtained is k, the inductance can be varied from $2L(1 - k)$ to $2L(1 + k)$. The construction of a typical variometer is shown in Fig. 25, where the two coils can be adjusted from series aiding to series opposing by rotating the inner coil, and where the maximum coefficient of coupling is made high by arranging the two inductances so that they coincide as nearly as is physically possible.

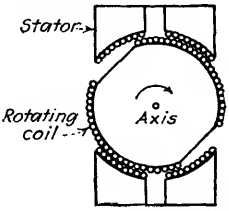


FIG. 25.—Constructional details of a typical variable inductance (variometer).

11. Electrostatic and Electromagnetic Shielding of Coils.—Under many conditions it is necessary to confine substantially all electrostatic and electromagnetic flux to a limited space in the neighborhood of the coil. This result can be accomplished by completely enclosing the

[1] Thus, where neither terminal is grounded and where there are more than two sections, it is desirable to make the self-inductance of the sections decrease from the ends toward the center, as shown in Fig. 24*b* and pointed out by Herman P. Miller, Jr., Multi-band R-F Choke Coil Design, *Electronics*, vol. 8, p. 254, August, 1935. Similarly with two section chokes, the desired type of behavior can be very closely realized by properly adjusting the spacing and relative inductance of the sections as outlined by Harold A. Wheeler, The Design of Radio-frequency Choke Coils, *Proc. I.R.E.* vol. 24, p. 850, June, 1936.

coil in a container made of material having low electrical resistivity, such as copper, aluminum, or brass, or by use of a conductor made of magnetic material. Such an arrangement acts as an electrostatic shield because it is a conductor and so constitutes a Faraday cage that screens the space external to the cage from electrostatic effects within. As far as electrostatic shielding is concerned, the exact nature of the shielding material is not highly important, and the shielding is substantially perfect if the container in which the coil is located is watertight or if its joints are lapped.

When the magnetic flux that is to be shielded is unidirectional or of audio frequency, shields composed of magnetic material of high permeability are employed. Such material acts as a magnetic short circuit which prevents the magnetic flux lines from extending to the space outside the container in which the coil is located and thus gives a shielding effect with magnetic flux that is analogous in all respects to the effect that a Faraday cage has on electrostatic flux. In order to obtain complete shielding it is necessary that the magnetic material of the shield be sufficiently thick to give a low reluctance to the flux entering the shield at one point and departing from it at another, so that there will be negligible magnetomotive force developed between different points on the shield. Otherwise some magnetic flux will not be intercepted by the shield. The best magnetic shields are those made of magnetic materials having high permeability at low flux densities, such as permalloy.

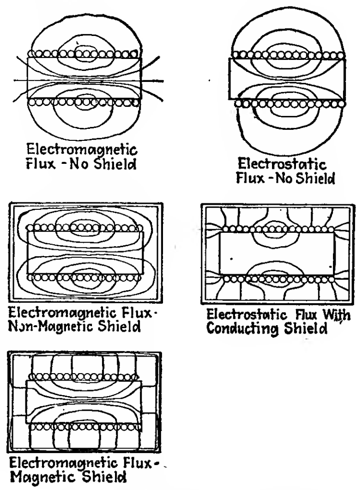


FIG. 26.—Paths of electrostatic and electromagnetic flux lines about the same coil with and without magnetic and non-magnetic shields.

Where the frequency of the magnetic field is high, more satisfactory shielding can be obtained by the use of shields having high electrical conductivity. The magnetic flux in attempting to pass through such a shield induces voltages that set up eddy currents which in large measure prevent the magnetic flux from penetrating through the shield. The shielding effect produced by eddy currents increases with frequency and with the conductivity of the shielding material. This is because the voltage induced by a flux line is proportional to frequency and because the eddy currents produced by a given induced voltage are proportional to the conductivity of the shield. At radio frequencies the result is that non-magnetic shields of low electrical resistance, such as copper and

aluminum, are better than high-resistance magnetic materials, such as iron. The effect that a non-magnetic shield has on the electrostatic and electromagnetic flux of a coil is shown in Fig. 26. Non-magnetic shields have no shielding effect on unidirectional fields.

When perfect electrostatic and magnetic shielding is required, it is necessary to employ several concentric shields separated by appreciable air gaps. These shields should be as nearly continuous electrically as is possible, and great care must be employed with leads that enter or leave the shields.[1]

Effect of Shielding on Coil Properties.—Placing a coil within a shield increases the coil's distributed capacity and the effective resistance, while the coil inductance is reduced with non-magnetic shields and increased when the shield is magnetic. The distributed capacity is increased as a result of the capacity between various parts of the coil and the shield. Restricting the magnetic-flux lines to the space within the shield by the use of non-magnetic material of high electrical conductivity increases the reluctance of the magnetic circuit (*i.e.*, decreases the amount of flux produced by a given coil current) and hence reduces the effective inductance of the coil by an amount that depends on the extent to which the shield interferes with the normal flux paths. When the shield is so large in comparison with the coil size as to interfere with only a small portion of the flux paths, the effect on the inductance will be small, whereas a close-fitting shield interferes with many flux lines and has the effect of greatly lowering the effective coil inductance. If the shield is of magnetic material, the effect on inductance is just the opposite, since the magnetic material supplies a low-reluctance path to the magnetic flux which increases the flux (and hence the inductance) the closer the shield is to the coil.

The energy consumed by the eddy currents flowing in the shield and by the iron losses if a magnetic shield is used must be supplied by the coil; this will have the effect of increasing the effective coil resistance by an amount that depends primarily upon the extent to which the shield interferes with the normal flux paths. If high-conductivity shielding material is used and if the shield is not too close to the coil, the added losses will not be excessive.

The shielding obtained by inclosing a coil in a copper or aluminum can is very good, and, if a reasonable clearance is provided between shield and coil, the properties of the coil are not seriously impaired. Thus with the usual coil, if the shield clears the coil everywhere by a distance at least equal to the coil radius, the coil inductance will not be

[1] A more complete discussion of the problems involved in obtaining complete shielding of radio-frequency fields is given by F. E. Terman, "Measurements in Radio Engineering," pp. 218–221, McGraw-Hill Book Company, Inc.

reduced by more than 10 to 15 per cent and the Q will be reduced even less.[1] A properly shielded single-layer or universal-wound coil will, in fact, have not only better shielding but also higher Q than a self-shielded coil, such as a toroid, of equal size.

Electrostatic Shielding without Magnetic Shielding.[2]—Electrostatic flux can be shielded without affecting magnetic fields by surrounding the space to be shielded with a conducting cage that is made in such a way as to provide no low-resistance path for the flow of eddy currents, at the same time offering a metallic terminal upon which electrostatic flux can end. One way of accomplishing this is to use a cage formed from a wire having one terminal grounded and the other free. Such a cage acts as a fairly good electrostatic shield but has very little influence on the magnetic field since there are no closed loops around which eddy currents can circulate. Another type of electrostatic shield commonly employed to shield primary and secondary windings in transformers consists of metal foil so arranged that its surface is approximately parallel with the magnetic field, with an insulated gap provided where necessary to prevent the shield from becoming a short-circuited turn.

12. Radio-frequency Coils with Magnetic Cores.—The use of magnetic cores in coils intended for radio-frequency service is limited by the fact that the eddy-current losses in magnetic material are proportional to the square of the frequency and so become excessively great when the frequency is high. The large eddy currents produced in the core also set up flux of their own which is in opposition to the main flux and therefore causes the effective inductance of the coil to become lower as the frequency is increased. This phenomenon is known as magnetic skin effect, since it is equivalent to preventing the magnetic flux from penetrating very deeply into the magnetic material of the core and so is analogous to skin effect in conductors.

The eddy currents can be reduced by subdividing the magnetic core material and by using magnetic material of high electrical resistivity. This can be accomplished in a fairly satisfactory manner at the lowest radio frequencies by the use of extremely thin laminations of magnetic materials such as silicon steel. A still better method of dividing the

[1] For further discussion of effect of shields upon coil properties, see G. W. O. Howe, The Effect of Screening Cans on the Effective Inductance and Resistance of Coils, *Wireless Eng. and Exp. Wireless*, vol. 11, p. 115, March, 1934; RCA Application Note 48, "Graphic Determination of the Decrease in Inductance Produced by a Coil Shield," obtainable from RCA Radiotron Corporation; W. G. Hayman, Inductance of Solenoids in Cylindrical Screen Boxes, *Wireless Eng. and Exp. Wireless*, vol. 11, p. 189, April, 1934.

Also see Sec. 17 for a qualitative analysis of the effect produced by a non-magnetic shield, based on coupled circuits.

[2] For further discussion see Terman, *op. cit.*, pp. 341–343, 345.

core consists of reducing the core material to a fine dust, mixing this powder with an insulating compound that will surround the particles, and then forming the mixture under high pressure into solid cores, preferably of toroidal shape. This construction produces a core that is subdivided to a much greater extent than can be realized with laminations, and at the same time there are distributed throughout its length air gaps that prevent the core from saturating or changing its inductance with moderate flux densities.

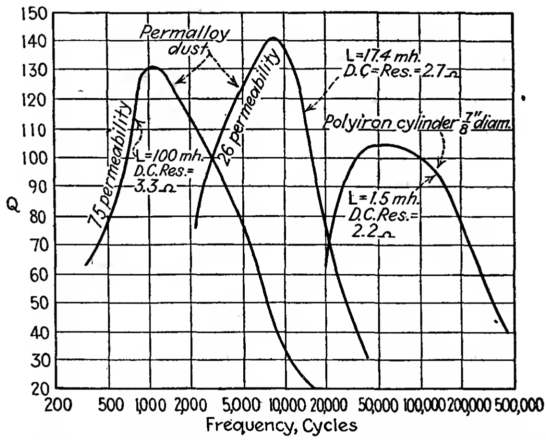


FIG. 27.—Curves showing values of Q as a function of frequency for various coils having dust cores.

Dust cores have been successfully used with coils intended for frequencies below about 500 to 1000 kc. The construction and design of the cores depends primarily upon the frequency range. In particular, coils intended for use at the high frequencies must be made from finer particles and the individual particles must be more thoroughly insulated from each other, which means a larger amount of insulating compound and a lower effective permeability when the resulting air gaps are taken into account. Iron- and permalloy-dust cores are used extensively in telephone work for audio and carrier frequencies.[1] These are constructed in the form of rings which have permeabilities ranging from 13 to 75, with the lower values obtained in rings designed for higher frequency service. More recently iron-dust cores suitable for use at intermediate

[1] A description of the method of preparation of cores made of permalloy dust and iron dust and their outstanding mechanical and electrical properties is to be found in the following articles: Buckner Speed and G. W. Elmen, Magnetic Properties of Compressed Powdered Iron, *Jour. A.I.E.E.*, vol. 40, p. 596, July, 1921; W. J. Shackelton and I. G. Barber, Compressed Powdered Permalloy, *Trans. A.I.E.E.*, vol. 47, p. 429, April, 1928.

and lower broadcast frequencies have been developed.[1] These cores are made from chemically produced iron-dust particles of about 5 μ diameter. The particles are insulated by a varnish, mixed with bakelite powder, and then molded under pressure to the final shape. The effective permeability obtained is of the order of 10 and can be controlled by varying the amount of bakelite. The cores are usually employed in the form of cylindrical slugs upon which a universal-wound coil is directly mounted. Such coils have been found to be extremely satisfactory at the intermediate frequencies used in broadcast receivers, and will have a fair Q even at the lower broadcast frequencies.

The performance of several types of dust-cored coils is given in Fig. 27. These results do not necessarily represent optimum designs, but they give an idea of what can be expected. It will be noted that in every case the Q drops off at high frequencies. This is because of increased eddy-current losses at the higher frequencies. The falling off in Q at low frequencies is caused by the increasing importance of the copper loss as the reduction in frequency reduces the eddy-current loss and the inductive reactance. The Q is maximum at the frequency where the best balance exists between copper and eddy losses.

Problems

1. If in Fig. 6 the flux shown is produced by a current of 0.1 amp., estimate the coil inductance. (Assume that Fig. 6 gives a two-dimensional representation of the flux lines in the three-dimensional coil.)

2. Design the winding of a coil to have 200 μh inductance, when the diameter is 2 in. and the ratio of length to diameter is 1.5. Use a standard-size wire.

3. Calculate the inductance of a coil consisting of 60 turns wound on a 3½-in. diameter cylinder with a pitch of 20 turns to the inch.

4. In iron-cored coils subjected to a direct-current magnetization it is found that, with a given direct current in the winding, there is an optimum-length air gap for which the incremental inductance at low alternating flux densities is maximum, and this air gap is greater the larger the direct current. Explain.

5. Assume that the coil of Prob. 3 is the secondary and that the primary is wound at one end of the secondary with 20 turns of fine wire having a pitch of 40 turns to the inch.

a. Calculate the inductance of the primary.

b. Calculate the mutual inductance between primary and secondary.

c. Calculate the coefficient of coupling.

6. A primary coil having an inductance of 100 μh is connected in series with a secondary coil of 240 μh, and the total inductance of the combination is measured as 146 μh.

a. Determine the mutual inductance.

b. Determine the coefficient of coupling.

c. Determine the inductance that would be observed if the terminals of one of the coils were reversed.

[1] See W. J. Polydoroff, Ferro-Inductors and Permeability Tuning, *Proc. I.R.E.*, vol. 21, p. 690, May, 1933.

7. A variable condenser with air dielectric is to have a total capacity of 350 $\mu\mu$f. If the spacing between plates is 0.07 cm, calculate the total area of dielectric required. If the rotating plates are semicircular with a radius of 3 cm, calculate the number of rotating plates required, assuming both sides of every plate are active.

8. A mica condenser with power factor 0.0004 has a capacity of 0.001 mf.

a. What are its equivalent series and shunt resistances at frequencies of 1000, 100,000, and 10,000,000 cycles?

b. What are its phase angles at these same frequencies?

9. The condenser of Prob. 8 is able to stand a direct-current potential of 5000 volts and is capable of dissipating safely 3 watts of heat. What is the highest frequency at which the heating will begin to limit the voltage rating, and what is the voltage rating at frequencies of 1, 100, 1000, and 10,000 kc?

10. A certain variable condenser having air dielectric with bakelite supports obtains 10 $\mu\mu$f of its capacity through the bakelite dielectric having a power factor of 4 per cent and the remainder of its capacity from the air, which has no losses. What is the equivalent-series resistance at 100 and 10,000 kc when the total capacity is 100 mmf (90 mmf from air and 10 mmf from bakelite)?

11. It is desired that a particular fixed condenser have the lowest possible inductance. Describe constructional features that will contribute to this result.

12. Explain why the distributed capacity of a coil is always increased by the wax or other coating used for protection against moisture.

13. Explain why moisture in the insulation is very detrimental to the properties of a coil.

14. Derive an equation giving the exact relation between the Q of a coil and the coil power factor, and from this calculate the error in the approximate relation: Power factor $= 1/Q$, when $Q = 50$.

15. In Fig. 18, calculate and plot: (a) the resistance as a function of frequency for the No. 20 wire case; (b) the ratio of alternating-current to direct-current resistance for the same coil.

16. Explain how the skin-effect phenomenon causes the true inductance of a coil to be slightly less at radio frequencies than at very low frequencies.

17. In transmitter inductances explain why copper tubing gives a resistance just as low as solid copper conductor of the same diameter, and why it is superior to conductor in the form of flat strip.

18. Transmitter inductances wound with bare copper tubing are sometimes plated to prevent corrosion and improve the appearance. Discuss the relationship between resistivity and thickness of the plated layer and the radio-frequency resistance of the coil.

19. It is observed in a non-magnetic shield can surrounding a solenoidal coil that the shielding is not affected appreciably by a joint in the shield provided this joint is in a plane perpendicular to the axis of the coil, but is very seriously reduced if the joint is in a plane that contains the axis of the coil. Explain.

CHAPTER III

PROPERTIES OF RESONANT CIRCUITS

13. Series Resonance.—When a constant voltage of varying frequency is applied to a circuit consisting of an inductance, capacity, and resistance, all in series, the current that flows depends upon frequency in the manner shown in Fig. 28. At low frequencies the capacitive reactance of the circuit is large and the inductive reactance is small, so that most of the voltage drop is across the condenser, while the current is small and leads the applied voltage by nearly 90°. At high frequencies the inductive reactance is large and the capacitive reactance low, resulting in a small current that lags nearly 90° behind the applied voltage, and most of the voltage drop is across the inductance. In between these two extremes there is a frequency, called the resonant frequency, at which the capacitive and inductive reactances are exactly equal and consequently neutralize each other, leaving only the resistance of the circuit to oppose the flow of current. The current at the resonant frequency is accordingly equal to the applied voltage divided by the circuit resistance, and is very large if the resistance is low.

The characteristics of a series resonant circuit depend primarily upon the ratio of inductive reactance ωL to circuit resistance R, *i.e.*, upon $\omega L/R$. This ratio is frequently denoted by the symbol Q and is called the circuit Q. In the usual resonant circuit the radio-frequency resistance of the circuit is made up almost solely of coil resistance because the losses in a properly constructed condenser are negligible in comparison with those of the coil. The result is that the circuit Q can ordinarily be taken as the Q of the coil alone, which was discussed in Sec. 9.

The general effect of different circuit resistances, *i.e.*, different values of Q, is shown in Fig. 28. It is seen that, when the frequency differs appreciably from the resonant frequency, *the actual current is practically independent of the circuit resistance, and is very nearly the current that would be obtained with no losses.* On the other hand the current at the resonant frequency is determined solely by the resistance. The effect of increasing the resistance of a series circuit is accordingly to flatten the resonance curve by reducing the current at resonance. This broadens the top of the curve, giving a more nearly uniform current over a band of frequencies near the resonant point, but does so by reducing the

selectivity of the tuned circuit (*i.e.*, the ability to discriminate between voltages of different frequencies).

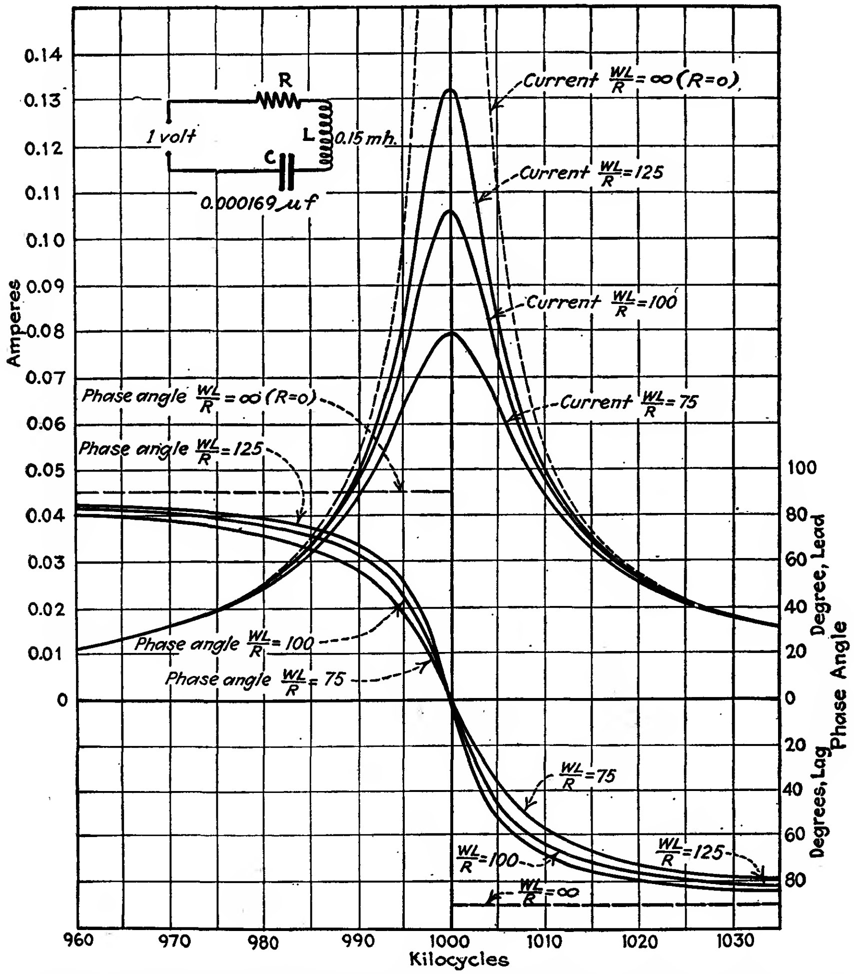


FIG. 28.—Magnitude and phase angle of current in series-resonant circuit as a function of frequency for a constant applied voltage. Curves are given for three values of circuit resistance and also for zero resistance.

Analysis of Series Resonant Circuit.—In deriving the relations that exist in a series resonant circuit, the following symbols will be used.

E = voltage applied to circuit
I = current flowing in circuit
f = frequency in cycles
$\omega = 2\pi f$

$Q = \omega L/R$
R = effective series resistance of tuned circuit
L = inductance in henrys
C = capacity in farads
Z = impedance of series circuit
θ = phase angle of impedance
Subscript $_o$ denotes values at resonant frequency

The resonant frequency is the frequency at which the inductive and capacitive reactances are equal; thus it is given by the expression

$$\omega L = \frac{1}{\omega C} \tag{21a}$$

or

$$\text{Resonant frequency} = f_o = \frac{1}{2\pi\sqrt{LC}} \tag{21b}$$

Equation (21*b*) shows that the resonant frequency depends only upon the product LC.

The impedance of the circuit is

$$Z = R + j\left(\omega L - \frac{1}{\omega C}\right) \tag{22}$$

and accordingly has a magnitude of

$$|Z| = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \tag{23a}$$

and a phase angle θ given by

$$\tan\theta = \frac{\left(\omega L - \frac{1}{\omega C}\right)}{R} \tag{23b}$$

The current flowing in the circuit is

$$I = \frac{E}{Z} = \frac{E}{R + j\left(\omega L - \frac{1}{\omega C}\right)} \tag{24a}$$

At resonance $Z = R$, giving a current in phase with the applied voltage and having a magnitude

$$\text{Current at resonance} = I_o = \frac{E}{R} \tag{24b}$$

This is the maximum current that can flow in the circuit.

In the series-resonant circuit the voltages across the condenser and the inductance will both be very much greater than the applied voltage when the frequency is in the vicinity of resonance. This is possible because the voltages across the condenser and inductance are nearly 180° out of phase with each other and so add up to a value that is much smaller than either voltage alone. Since the current at resonance is E/R, the voltage across the inductance at resonance is ωL times the current, or

$$\text{Voltage across } L \text{ at resonance} = E\frac{\omega L}{R} = EQ \tag{25}$$

The voltage across the condenser also has this same value since, at resonance, $\omega L = 1/\omega C$. Equation (25) shows that *at resonance the voltage across the inductance (or condenser) is Q times the applied voltage* (i.e., *there is a resonant rise of voltage in the circuit amounting to Q times*). Since Q can be expected to have a value in the neighborhood of 100, a series-resonant circuit will develop a high voltage even with small applied potentials. Thus, if the applied voltage is 50 and $Q = 100$, the potential developed across the condenser is 5000 volts and, unless the condenser is built for high-voltage service, the insulation will break down.

Universal Resonance Curve.—Equations (24*a*) and (24*b*) can also be rearranged to express the ratio of current actually flowing to the current at resonance, in terms of the circuit Q and the fractional deviation of the frequency from resonance. When this is done the result is[1]

[1] The derivation of this equation follows: The ratio of Eq. (24*a*) to (24*b*) gives:

$$\frac{\text{Actual current}}{\text{Current at resonance}} = \frac{R_o}{R + j\left(\omega L - \dfrac{1}{\omega C}\right)}$$

$$= \frac{R_o}{R + j\left(\dfrac{\omega^2 LC - 1}{\omega C}\right)}$$

By the definition of δ,

$$\omega = \omega_o(1 + \delta)$$

Substituting this value of ω and remembering that $\omega_o L = 1/\omega_o C$, one obtains

$$\frac{\text{Actual current}}{\text{Current at resonance}} = \frac{R_o}{R + j\left[\dfrac{(1+\delta)^2 - 1}{1+\delta}\right]\omega_o L}$$

$$= \frac{1}{\dfrac{R}{R_o} + jQ\delta\left(\dfrac{2+\delta}{1+\delta}\right)}$$

When Q is constant, the radio-frequency resistance is proportional to frequency so that $R/R_o = \omega/\omega_o = (1 + \delta)$, which when substituted yields Eq. (26).

$$\frac{\text{Actual current}}{\text{Current at resonance}} = \frac{1}{1 + \delta + jQ\delta\left(\frac{2 + \delta}{1 + \delta}\right)} \tag{26}$$

where

$$Q = \omega_o L/R_o = \text{circuit } Q \text{ at resonant frequency}$$
$$\delta = \frac{(\text{actual frequency}) - (\text{resonant frequency})}{(\text{resonant frequency})}$$

The quantity δ represents the ratio of the number of cycles by which the actual frequency is off resonance to the resonant frequency; this can be called the fractional deviation of the frequency from resonance. Thus a value $\delta = 0.01$ means that the actual frequency differs from the resonant frequency by 0.01 of the resonant frequency. The value of δ is positive or negative as the actual frequency is greater than or less than resonance, respectively. By expressing δ as $\delta = a/Q$, it is possible to plot the *universal resonance curves of Fig. 29, from which the exact resonance curve of any series circuit may be obtained without calculation when Q is known.* These curves are extremely useful because they are independent of the resonant frequency of the circuit and of the ratio of inductance to capacity. It is to be noted that they are practically symmetrical about the resonant frequency and are substantially independent of Q.

The use of Fig. 29 in practical calculations can be illustrated by two examples.

Example 1.—It is desired to know how many cycles one must be off resonance to reduce the current to one-half the value at resonance when the circuit has a Q of 125 and is resonant at 1000 kc.

Reference to Fig. 29 shows that the response is reduced to 0.5 when $a = 0.86$. Hence

$$\text{Cycles off resonance} = \frac{0.86 \times 1000}{125} = 6.88 \text{ kc}$$

The phase angle of the current as obtained from the curve is 60°.

Example 2.—With the same circuit as in the preceding example, it is desired to know what the response will be at a frequency 10,000 cycles below resonance.

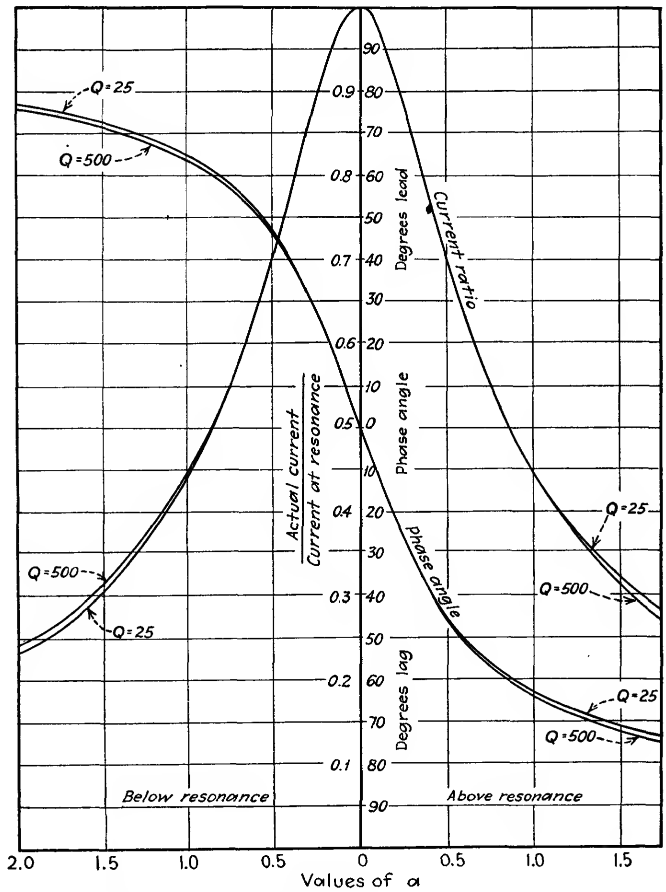
To solve this problem it is first necessary to determine a.

$$a = {}^{10}\!/_{1000} \times 125 = 1.25$$

Reference to Fig. 29 shows that for $a = 1.25$ the response is reduced by a factor 0.36, and that the phase of the current is 68° leading.

The only assumption involved in Eq. (26) is that Q is the same at the frequency being considered as at the resonant frequency. *When this is true, Eq. (26) and the universal resonance curve involve no approximations whatsoever.* Over the limited range of frequencies near resonance,

where the universal resonance curve finds its chief usefulness, the variations in Q are so small as to introduce negligible (*i.e.*, less than 1 per cent)



$$\frac{\text{Cycles off resonance}}{\text{Resonant frequency}} = \frac{a}{Q} = \text{Fractional detuning}$$

$$\text{Cycles off resonance} = \frac{a}{Q} \times \text{resonant frequency}$$

$$a = Q\,\frac{\text{Cycles off resonance}}{\text{Resonant frequency}}$$

FIG. 29.—Universal resonance curve from which the exact ratio of actual current to current at resonance, as well as exact phase angle, can be determined for any series circuit in terms of the detuning from resonance. This curve can also be applied to the parallel resonant circuit by considering the vertical scale to represent the ratio of actual parallel impedance to the parallel impedance at resonance. When applied to parallel circuits, angles shown in the figure as leading are lagging, and *vice versa*.

errors from the use of the curve provided the value of Q existing at resonance is used in determining the fractional deviation from resonance.

The voltage developed across the inductive or capacitive reactance of a series-resonant circuit is equal to the product of the current and the reactance. For most practical purposes the voltage developed across the inductance or capacity can be considered as varying with frequency in exactly the same way as does the current in the circuit. This is because the part of the resonance curve in which one is normally interested occupies a very narrow frequency band extending only a few per cent on either side of resonance, and for this limited range of frequencies the reactance of the inductance or capacity is substantially constant.

Working Rules for Estimating Sharpness of Resonance.—Since the curves for different values of Q are almost identical in Fig. 29, particularly in the neighborhood of the resonant frequency, it is possible to state several easily remembered working rules that will enable one to estimate the sharpness of any resonance curve with an error of less than 1 per cent when only the Q of the circuit is known.[1] These rules follow:

Rule 1. *When the frequency of the applied voltage deviates from the resonant frequency by an amount that is* $1/2Q$ *of the resonant frequency, the current that flows is reduced to* 70.7 *per cent of the resonant current, and the current is* 45° *out of phase with the applied voltage.*

Rule 2. *When the frequency of the applied voltage deviates from the resonant frequency by an amount that is* $1/Q$ *of the resonant frequency, the current that flows is reduced to* 44.7 *per cent of the resonant current, and the current is* $63\frac{1}{2}°$ *out of phase with the applied voltage.*

Thus in the circuit considered in the above examples the current would be reduced to 70.7 per cent of the value at resonance when the frequency is $\frac{1}{250}$ of 1000 kc, or 4000 cycles off resonance, and to 44.7 per cent of the resonant current for a frequency deviation of $\frac{1}{125}$ of 1000 kc, or 8000 cycles. Since the resonant rise of voltage in this circuit is 125 times, the rise of voltage is very nearly $0.7 \times 125 = 87.5$ times when the frequency is 4000 cycles off resonance, and is very close to $0.45 \times 125 = 56.25$ times at a frequency 8000 cycles from resonance.

Practical Calculation of Resonance Curves.—When a complete resonance curve is to be determined, it is necessary to supplement the above working rules by additional calculated points. The proper procedure is to start by calculating the current at resonance, using Eq. (24*b*). The working rules can then be applied to obtain the response at frequencies $1/2Q$ and $1/Q$ on either side of resonance. This gives a picture of the resonance curve and is sufficient for many purposes. Additional points as needed in the vicinity of resonance can be calculated with the

[1] An error of 1 per cent is nearly always permissible in calculations of radio-frequency circuits. This is because the effective circuit constants at radio frequencies are very seldom known to an accuracy that involves an error of less than 1 per cent.

aid of the universal resonance curve of Fig. 29. Currents at frequencies too far off resonance to come within the range of the universal resonance curve can be obtained with an accuracy sufficient for all practical purposes by neglecting the resistance. This introduces an error of only 1 per cent when the number of cycles off resonance is equal to the resonant frequency multiplied by $3.5/Q$, with rapidly diminishing error still farther off resonance. When this simplification is made,[1]

$$\frac{\text{Actual current when far off resonance}}{\text{Current of resonance}} = \frac{1}{Q\gamma\left(1 - \frac{1}{\gamma^2}\right)} \tag{27a}$$

where

$$\gamma = \frac{\text{actual frequency}}{\text{resonant frequency}}$$

The phase angle of the current is usually not important, but if desired it can be calculated without error by the use of Eq. (23b), which can also be rewritten

$$\tan \theta = Q\left(1 - \frac{1}{\gamma^2}\right) \tag{27b}$$

The above procedure for calculating resonance curves is much superior to making calculations based directly upon Eq. (24a). The use of the universal resonance curve in the vicinity of the resonant frequency not only reduces the amount of labor involved but also greatly improves the accuracy under ordinary conditions. This is because resonant circuit formulas such as Eq. (24a) contain a term $\left(\omega L - \frac{1}{\omega C}\right)$ which involves the difference of two large and nearly equal quantities. In order to obtain this difference without more than 1 per cent error, five-place logarithms must ordinarily be employed. Slide-rule calculations are never per-

[1] Equation (27a) is derived as follows: When resistance is neglected, Eq. (24a) becomes

$$I = \frac{E}{j\left(\omega L - \frac{1}{\omega C}\right)} = \frac{E}{j\omega L\left(1 - \frac{1}{\omega^2 LC}\right)}$$

Dividing both numerator and denominator by R_0, the resistance at resonance, and substituting $\omega = \gamma\omega_0$ gives

$$|I| = \frac{E/R_0}{\frac{\gamma\omega_0 L}{R_0}\left(1 - \frac{1}{\gamma^2\omega_0^2 LC}\right)}$$

Equation (27a) follows when it is remembered that $1/\omega_0^2 LC = 1$ because at resonance $\omega_0 L = 1/\omega_0 C$.

missible. Neglecting the resistance at frequencies too far off resonance to come within the range of the universal resonance curve reduces the labor enormously and introduces an error of less than 1 per cent of the magnitude at resonance. This accuracy is ample for all ordinary purposes and the error is undetectable when resonance curves are plotted.

The resonance curve of voltage developed across the inductance or capacity of a series circuit is obtained in much the same way as the current. For precise results one calculates the current resonance curve and then multiplies the result by the reactance across which the voltage is developed. However, for ordinary purposes the recommended procedure is to determine the voltage at resonance, which has already been shown to be Q times the applied voltage. The universal resonance curve is then applied to determine the way this voltage drops off on either side of resonance for frequencies in the vicinity of resonance. For frequencies too far off resonance to come within the range of the universal resonance curve, calculations of magnitude can be made satisfactorily by neglecting the resistance. When this is done, a little manipulation based on Eq. (24*a*) gives:

For voltage across the inductance:

$$\frac{\text{Actual voltage when far off resonance}}{\text{Voltage at resonance}} = \frac{1}{Q\left(1 - \dfrac{1}{\gamma^2}\right)} \qquad (28a)$$

For voltage across the condenser:

$$\frac{\text{Actual voltage when far off resonance}}{\text{Voltage at resonance}} = \frac{1}{Q(\gamma^2 - 1)} \qquad (28b)$$

where Q and γ have the values as defined above.

14. Parallel Resonance.—A parallel circuit consisting of an inductance branch in parallel with a capacity branch offers an impedance of the character shown in Fig. 30. At very low frequencies the inductive branch draws a large lagging current while the leading current of the capacity branch is small, resulting in a large lagging line current and a low lagging circuit impedance. At high frequencies the inductance has a high reactance compared with the capacity, resulting in a large leading line current and a correspondingly low circuit impedance that is leading in phase. In between these two extremes there is a frequency at which the lagging current taken by the inductive branch and the leading current entering the capacity branch are equal; and being 180° out of phase, they neutralize, leaving only a small resultant in-phase current flowing

in the line. The impedance of the parallel circuit will then be a very high resistance, as is brought out in Fig. 30.[1]

The effect of circuit resistance upon the impedance of the parallel-resonant circuit is very similar to the influence that resistance has

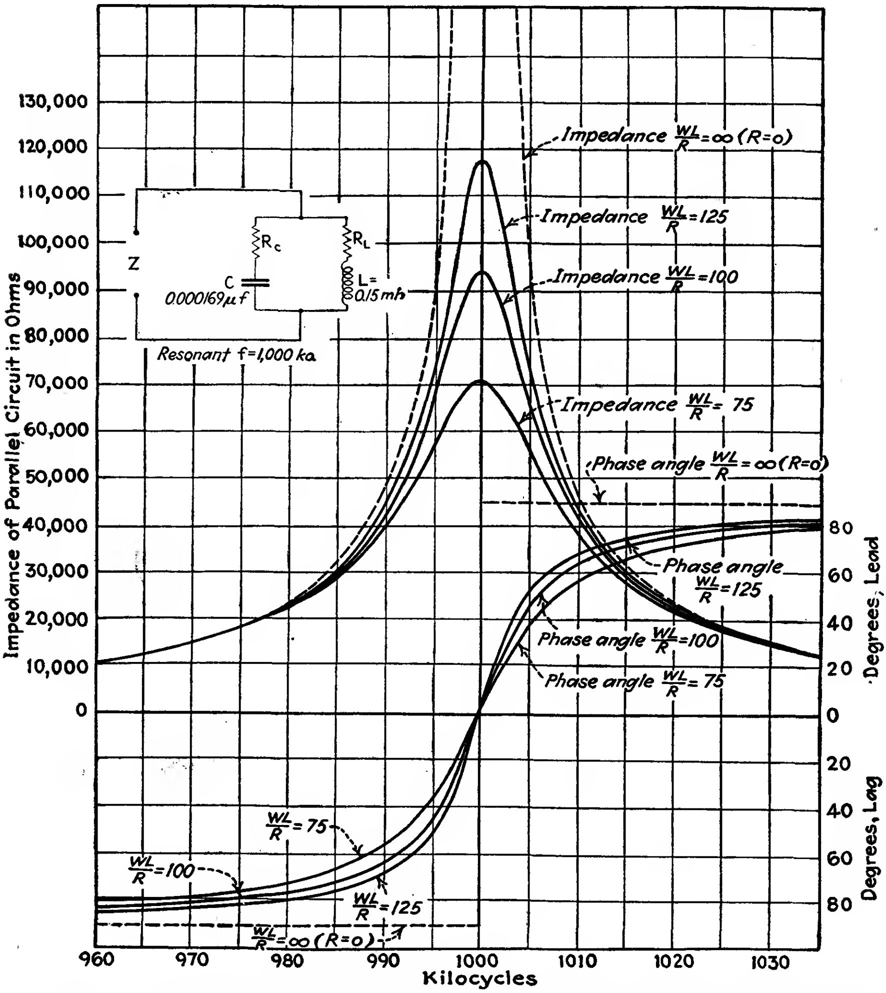


FIG. 30.—Magnitude and phase angle of impedance of a parallel circuit as a function of frequency. Curves are given for three representative values of circuit resistance and also for zero resistance.

upon the current flowing in a series-resonant circuit, as is evident when Figs. 28 and 30 are compared. Increasing the resistance of a parallel

[1] In obtaining a parallel-resonance curve experimentally by measurements of applied voltage and line current, extreme care must be taken to insure that the applied voltage contains no harmonics. This is necessary because at resonance the

circuit lowers and flattens the peak of the impedance curve without appreciably altering the sides, which are relatively independent of the circuit resistance. The dotted curve of Fig. 30 represents the impedance of the parallel circuit as a function of frequency when the circuit has no losses. It is apparent that at frequencies that differ from the resonant frequency by at least several per cent the resistance has little effect on the magnitude of the impedance.

The resonant frequency of the parallel circuit is sometimes taken as the point of minimum line current, sometimes as the condition that makes the impedance a pure resistance, and sometimes as the frequency for which $\omega L = 1/\omega C$. These three definitions of resonance in parallel circuits lead to resonant frequencies that are different by such a very small fraction of 1 per cent when the circuit Q is at all large that for all practical purposes the resonant frequency of a parallel circuit can be taken as the frequency that satisfies the relation

$$\omega L = \frac{1}{\omega C} \tag{29a}$$

or

$$\text{Resonant frequency} = \frac{1}{2\pi\sqrt{LC}} \tag{29b}$$

With this definition the parallel resonant frequency of a tuned circuit is exactly the same as the series-resonant frequency of the same circuit.

Analysis of Parallel Resonance.—In the analysis of the parallel-resonant circuit of Fig. 30 the following notation will be used:

$$\begin{aligned}
E &= \text{voltage applied to circuit}\\
Z_c &= R_c - \frac{j}{\omega C} = \text{impedance of capacitive branch}\\
Z_L &= R_L + j\omega L = \text{impedance of inductive branch}\\
Z_s &= Z_c + Z_L = \text{series impedance of circuit}\\
Z &= \text{parallel impedance of circuit}\\
R_s &= R_c + R_L = \text{total series resistance of circuit}\\
\omega &= 2\pi \text{ times frequency}\\
\omega_0 &= 2\pi \text{ times resonant frequency}\\
Q &= \frac{\omega L}{R_s}
\end{aligned}$$

The quantities L, C, R_L, and R_c refer to the inductance, capacity, and resistance components of the circuit as indicated in Fig. 30.

circuit impedance is extremely high to the fundamental component of the applied voltage and very low to the harmonic components, with the result that even a small harmonic-voltage component will cause line currents that mask the small fundamental component.

The impedance of the parallel circuit is the impedance formed by the capacitive and inductive branches connected in parallel:

$$\text{Parallel impedance} = Z = \frac{Z_c Z_L}{Z_c + Z_L} = \frac{Z_c Z_L}{Z_s} \tag{30a}$$

It will be noted that the denominator of this expression is the impedance that the same circuit offers when connected in series, as given by Eq. (22). Equation (30a) is the fundamental equation of the parallel circuit and applies for all conditions, irrespective of the circuit Q, the frequency, or the division of resistance between the branches.

When the circuit Q is reasonably high, as is nearly always the case in actual practice, the exact expression of Eq. (30a) can be simplified by neglecting the resistance components of the impedances Z_L and Z_C in the numerator. When this is done,[1]

$$\text{Parallel impedance} = Z = \frac{(\omega_0 L)^2}{Z_s} \tag{30b}$$

The error involved in Eq. (30b) depends upon the Q of the circuit and the division of resistance between the branches. Under the most unfavorable condition, which is when all the resistance is in one branch, the magnitude of the parallel impedance as calculated by Eq. (30b) has an error of one part in $2Q^2$, while the error in phase angle is $\tan^{-1}\frac{1}{Q}$. These errors are entirely negligible for ordinary calculations if the circuit Q is at all high. Thus for $Q = 100$ the error in magnitude is $\frac{1}{200}$ of 1 per cent, and the error in phase is 0.5°.

Examination of Eq. (30b) shows that the parallel impedance is equal to a constant divided by the series impedance of the circuit. This means that the resonance curve of parallel impedance of a circuit has exactly the same shape as the resonance curve of the series current in the same circuit, since in Eq. (24a) the current of a series circuit is a constant divided by the series impedance. Consequently the universal resonance curve and the working rules that were applied for estimating the sharpness of resonance of the series circuit also apply to the case of parallel resonance. The only difference is that the signs of the phase angles are now reversed, the phase now being leading at frequencies higher than resonance and lagging at frequencies below resonance.

[1] This transformation is carried out as follows: If the resistance components are neglected, the product $Z_L Z_C$ becomes $\omega L/\omega C = L/C$. One can now eliminate the capacity C in this expression by multiplying both numerator and denominator by ω_0 and then noting that $1/\omega_0 C = \omega_0 L$. That is,

$$Z_L Z_C = \frac{L}{C} = \frac{\omega_0 L}{\omega_0 C} = (\omega_0 L)^2$$

Calculation of Parallel Impedance.—The proper procedure for calculating the impedance of a parallel-resonant circuit is similar to that used with a series circuit. The first step is the determination of the impedance at resonance. At the resonant frequency the series impedance Z_s becomes the series resistance R_s of the circuit, so that Eq. (30*b*) becomes

$$\text{Parallel impedance at resonance} = \frac{(\omega_0 L)^2}{R_s} \tag{31a}$$

$$= (\omega_0 L)Q \tag{31b}$$

The next step is to apply the working rules to obtain the 70.7 per cent and 44.7 per cent points on either side of resonance. This gives the general picture of the sharpness of resonance, and is sufficient for many purposes. If a more complete resonance curve is desired, however, one can make use of the universal resonance curve of Fig. 30 to calculate additional points in the region around resonance. Finally, points too far off resonance to come within the range of the universal curve can be calculated with an accuracy sufficient for all ordinary purposes by neglecting the circuit resistance. With this simplification, Eq. (30*b*) becomes

$$\frac{\text{Impedance at frequencies far from resonance}}{\text{Impedance at resonance}} = \frac{1}{Q\gamma\left(1 - \dfrac{1}{\gamma^2}\right)} \tag{32a}$$

where γ is the ratio of the actual to the resonant frequency. The phase angle at all frequencies is very nearly the negative of the angle of the current for series operation, and from Eq. (27*b*) is found to be

$$\tan \theta = Q\left(\frac{1}{\gamma^2} - 1\right) \tag{32b}$$

Miscellaneous Features of Parallel Resonance.—It will be noted from Eq. (31*a*) and (31*b*) that at resonance the impedance of the parallel circuit is a resistance that is Q times the impedance of one of the branches. It can therefore be said that the parallel arrangement of inductive and capacitive branches causes a resonant rise of impedance of Q times the impedance that would be obtained from either branch alone. It is thus apparent that very high impedances can be developed by parallel resonance, and this is one of the most important properties of parallel resonance.

The line current that flows when a voltage is applied as in Fig. 30 is relatively small at the resonant frequency because of the high circuit impedance, but, as the frequency departs from resonance, becomes rapidly larger as the result of the lowered circuit impedance. The currents in the individual branches do not go through any resonance

action, however, since the current in a branch is equal to the applied voltage divided by the branch impedance, and this impedance varies only very slowly with frequency. The resonance phenomenon arises from the fact that at the resonant frequency both capacitive and inductive branches draw large currents, but these currents are reactive and add up to a very small resultant. This means that at resonance there is a large current circulating between the inductance and capacity, with the line current being just large enough to supply the circuit losses. Inasmuch as the resistance of a tuned circuit is low, the energy losses and hence the line current will be correspondingly small at resonance. At frequencies off resonance, the line current increases because the reactive currents drawn by the capacitive and inductive branches are not equal, which makes it necessary for the line to supply considerable reactive current along with the power component of current.

The quantitative relations between line and branch currents follow:

$$\text{Line current} = \frac{E}{Z} \tag{33a}$$

$$\text{Inductive branch current} = \frac{E}{Z_L} = \frac{E}{R_L + j\omega L} \tag{33b}$$

$$\text{Capacitive branch current} = \frac{E}{Z_C} = \frac{E}{R_C - \dfrac{j}{\omega C}} \tag{33c}$$

At resonance the two branch currents are substantially the same, and have a value that is very nearly $E/\omega_0 L$, while the line current is $E/(\omega_0 L)Q$. *At resonance the branch currents are Q times the line current, giving what might be thought of as a resonant rise of current from line to branch of Q times.*

When the voltage is applied across only part of the inductive or part of the capacitive branch of a parallel resonant circuit, as illustrated in Fig. 31, the only effect on the impedance near the resonant frequency is to reduce the magnitude of the parallel impedance curve without changing its shape. The magnitude actually obtained will be approximately proportional to the square of the branch reactance between the points of voltage application. Thus the curve of parallel impedance for the case in Fig. 31*a* when the voltage is applied across only one-half of the coil will have exactly the same shape as though the voltage has been applied across the entire inductance, but the impedance values will be only one-fourth as great. Tapping in at intermediate points on a branch of a parallel circuit accordingly serves to step down the parallel impedance to lower values without changing its characteristics near the parallel resonant frequency.

Parallel Circuits with Low Q.—When the circuit Q is low, the approximations made in deriving Eq. (30*b*) will introduce an error that depends on the division of resistance between the inductive and capacitive branches. If the resistances are divided so that $\omega L/R_L$ and $1/\omega CR_c$ are approximately equal, the error is at a minimum. The parallel impedance then very accurately follows the behavior discussed above. If, on the other hand, the resistances in the two branches are decidedly unequal, as is commonly the case since the losses are usually concentrated in the inductive branch, then Eq. (30*b*) and the discussion based upon it are only approximately true. The principal modification produced under such conditions is that the frequency at which the parallel circuit has maximum impedance is not the frequency for which it has unity power factor. This behavior begins to appear when the circuit Q drops below 10 with the resistance largely concentrated in one branch. The exact behavior for low circuit Q's can be calculated from Eq. (30*a*) by using the exact expressions for Z_L and Z_C in determining the numerator. The division of resistance between the inductive and capacitive branches of a parallel circuit has relatively little effect on the circuit behavior provided the Q is not too low. The most important thing is then the value of total resistance R_s and the resulting value of $\omega L/R_s$.

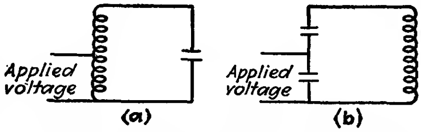


FIG. 31.—Parallel resonant circuits connected to make the parallel impedance at the circuit terminals less than the impedance across the complete circuit.

Components of Parallel Impedance.—The parallel impedance as calculated by Eq. (30*a*) can be thought of as equivalent to a resistance in series with a reactance. These resistance and reactance components will be found to vary with frequency in the manner shown in Fig. 32. The resistance component varies in much the same way with frequency as does the parallel impedance, and it reaches a magnitude at resonance equal to the parallel impedance. The sides of the resistance curve fall away faster than does the resonance curve, however. The reactance of the parallel circuit is inductive at low frequencies, reaches a maximum value at a frequency slightly below resonance, passes through zero at resonance, and then becomes a capacitive reactance at higher frequencies. The maximum values of inductive and capacitive reactance are both almost exactly one-half the impedance of the parallel circuit at resonance, and the reactance maxima occur at a number of cycles off resonance equal to the resonance frequency divided by $2Q$.

Parallel Resonance Effects in Inductance Coils.—The self-capacity associated with an inductance coil is in shunt with the inductance and thus makes the coil equivalent to a parallel-resonant circuit. The result is that the apparent coil inductance as measured between the

terminals increases with frequency until a maximum is reached just below the frequency at which the self-capacity is resonant with the coil inductance. The apparent inductance becomes zero at the parallel-resonant frequency, *while for higher frequencies the coil has a capacitive reactance and is therefore equivalent to a small condenser.* The apparent resistance

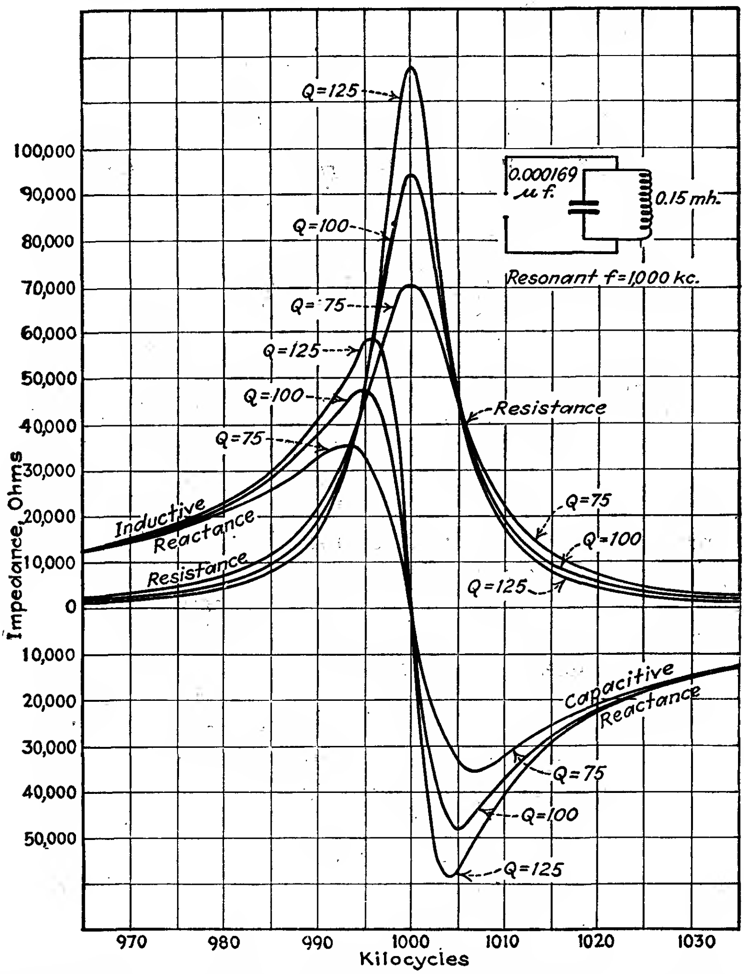


FIG. 32.—Resistance and reactance components of the parallel impedance of Fig. 30 shown as a function of frequency.

of the coil increases with the frequency until a maximum is reached at the resonant frequency, beyond which the resistance rapidly diminishes. These effects are all direct consequences of the properties of parallel-resonant circuits and can be readily deduced by an examination of Fig. 32 or of Eqs. (30*a*) and (30*b*).

The behavior of an inductance coil with self-capacity can be calculated just as one would determine the characteristics of any parallel circuit. The frequency at which the self-capacity and coil inductance are in parallel resonance is called the natural frequency of the inductance coil and represents the highest frequency at which the coil acts as an inductance. At frequencies that do not exceed 80 per cent of the natural frequency, Eq. (30a) can be rearranged to yield the following approximate results:[1]

$$\text{Apparent inductance of coil with self-capacity} = \frac{L}{1 - \gamma^2} \tag{34}$$

$$\text{Apparent resistance of coil with self-capacity} = \frac{R}{(1 - \gamma^2)^2} \tag{35}$$

where γ is the ratio of actual frequency to natural resonant frequency of the coil while L and R are the true coil inductance and resistance, respectively. The apparent inductance and resistance given by these equations are the values one would find by actual measurement and they represent the apparent but not the true quantities. Thus at a frequency one-half the natural frequency of the coil, $\gamma = \frac{1}{2}$ and $(1 - \gamma^2) = \frac{3}{4}$, and as a result of the coil capacity the apparent inductance and resistance as measured are, respectively, $\frac{4}{3}$ and $\frac{16}{9}$ of the actual values. It is to be noted that the apparent Q of a coil is decreased by self-capacity. Because of the parallel-resonance effects resulting from self-capacity, it is customary to determine the inductance of radio coils at audio frequencies, where γ is very small.

This discussion of the effect of self-capacity in inductance coils is approximate in that it assumes the self-capacity is lumped across the terminals of the coil as a single condenser, whereas it is actually distributed throughout the coil as shown in Fig. 14. The principal effect of assuming the self-capacity lumped instead of distributed is to make the calculated natural frequency in error, but for frequencies that are appreciably lower than the natural period the approximation is very good.

15. Voltage and Current Distribution in Circuits with Distributed Constants.[2]—When the inductance, capacity, and resistance of a circuit

[1] Equations (34) and (35) are derived by rationalizing the denominator of Eq. (30a) and then neglecting the resistance term of the denominator as being small compared with the terms involving only the inductance and capacity when the frequency is not too close to resonance. The real part of the result is the apparent resistance given in Eq. (35), and the reactive part can be simplified to give Eq. (34).

[2] The subject of circuits with distributed constants is so extensive as to require an entire book devoted to this one topic if the treatment is to be adequate. The remarks made in this section should be considered as giving only a superficial treatment of those properties of such circuits which are of fundamental importance in radio communication. For further information relative to circuits with distributed constants, the reader is referred to any good book on the theory of telephone or power lines.

are mixed together rather than being in separate lumps as in the case of the simple series and parallel circuits that have been considered, it is said that the circuit has distributed constants. Examples of circuits with distributed constants include telephone, telegraph, and power lines, as well as most types of radio antennas.

The distribution of voltage and current along a circuit with distributed constants, such as a transmission line, depends upon the impedance at the receiving end of the circuit and upon the distance from the receiving end to the point at which the voltage is applied to the circuit. The distributions with open- and short-circuited receivers are shown in Fig. 33 and are seen to involve very pronounced resonances. In considering these resonances it is convenient to measure distances along the circuit in terms of wave lengths. One wave length is twice the distance between adjacent minima and is given with high degree of accuracy by the equation

$$\left.\begin{array}{l}\text{Distance along circuit corre-}\\ \text{sponding to one wave length}\end{array}\right\} = \frac{1}{f\sqrt{LC}} \tag{36}$$

in which L and C are the series inductance and shunt capacity, respectively, per unit length of circuit, and f is the frequency of the applied voltage. Where the circuit consists of one or more straight wires in space, as is the case with an antenna or a 60-cycle power line, the distance corresponding to one wave length found from Eq. (36) will always be almost exactly the same as the wave length of radio waves of the same frequency.

Voltages and Current Distribution with Open- and Short-circuited Receiver.—When the receiving end of a circuit with distributed constants is open, the voltage distribution is such that the voltage goes through minima at distances from the receiver corresponding to an odd number of quarter wave lengths and goes through maxima at distances corresponding to an even number of quarter wave lengths, always measured from the receiver. The current distribution for the open-circuited receiver has its minimum values where the voltage is maximum and its maximum values where the voltage is smallest. The ratio of these maximum to minimum values in a distribution curve, such as Fig. 33*a*, depends on the resistance per unit length of circuit and is greater the lower the resistance. Changing the frequency alters the distance representing one wave length and so changes the number of maxima and minima obtained in a given line length but does not otherwise affect the general character of the line behavior. The sending-end impedance of a circuit with distributed constants having an open receiver will be low when the line is an odd number of quarter wave lengths long because then the sending-end voltage is small and the current is high, while the impedance

will be high when the line is an even number of quarter wave lengths in length.

The voltage that must be applied at the sending end of a circuit with distributed constants to develop a given receiving-end voltage (E_r in Fig. 33*a*) on open circuit depends upon the length of the line in wave lengths. The ratio of receiver voltage to sending-end voltage is equal to the ratio of E_r in Fig. 33*a* to the height of the voltage curve in Fig. 33*a* at a distance from the receiver corresponding to the length of the line. Thus, if the line is a quarter wave length long, the sending end is at *A*

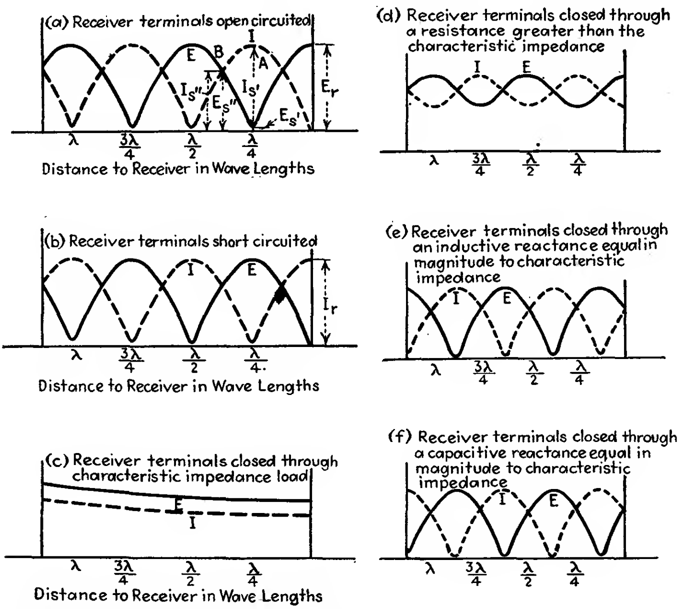


FIG. 33.—Typical voltage and current distributions for circuit with distributed constants. showing conditions existing with widely different receiver end conditions.

and an applied voltage E_s' will produce an open-circuit receiver voltage E_r; if the line is three-eighths wave lengths long, the sending end is at *B* and an applied voltage E_s'' will be necessary to give a receiver voltage of E_r. The currents that will flow into two such lines will be I_s' and I_s'', respectively, as shown in the figure. It is seen, therefore, that, when the circuit is exactly some odd number of quarter wave lengths in length, there will be a resonant rise of voltage in the line, while, if it is exactly an even number of quarter wave lengths long, there will be no place along the line where the voltage exceeds the sending-end potential.

When the receiving end of a circuit with distributed constants is short-circuited, the voltage and current distribution are as shown in Fig. 33*b*. It will be noted that the voltage distribution in the short-circuited case is

of exactly the same character as the current in the open-circuited receiver case, and the current distribution with the receiver shorted is exactly the same as the voltage distribution with open receiver terminals. By making allowances for this interchange of voltage and current, everything that was said about the open-circuited receiver also applies to the case of the short-circuited receiver. Thus, when the receiving terminals are short-circuited, the voltage along the line is low at points that are an even number of quarter wave lengths from the receiver and the current is low at all odd quarter wave lengths from the receiver. The sending-end impedance is low when the line length is an even number of quarter wave lengths long and is high at lengths that are an odd number of quarter wave lengths distant from receiver.

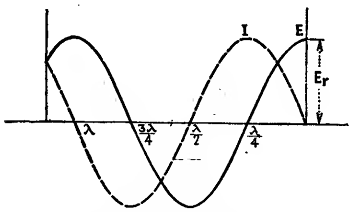


FIG. 34.—Schematic method of representing voltage and current distribution of Case *a* in Fig. 33. This diagram indicates the 180° phase shift on opposite sides of the minima and is an accurate representation of the conditions existing along the circuit except where the voltage and current are shown going through zero.

In the cases of both open- and short-circuited receivers the voltage and current are substantially 90° out of phase at all places along the line except in the vicinity of the quarter-wave-length points, where the phase angle rapidly shifts from nearly 90° on one side of unity power factor to nearly 90° on the other side of unity power factor. The voltages on opposite sides of a voltage minimum are therefore substantially 180° out of phase, as are also the currents on opposite sides of a current minimum. In order to show this change of phase, the voltage and current distributions in circuits with distributed constants are frequently drawn as shown in Fig. 34, which corresponds to *a* of Fig. 33 except that adjacent maxima are shown on opposite sides of the base line to indicate the reversal of phase. It will be noted that, while this method of representation gives the distribution along most of the line accurately, it fails to show the conditions actually existing at the minima since the curves of Fig. 34 go through zero where the axis is crossed, whereas in reality the minima never reach zero if the line has any losses, and where there are losses, the 180° phase shift takes place gradually, instead of all at one spot. In many circumstances, notably in the calculation of radiation from a circuit with distributed constants, the small amplitude at the minima is of negligible importance and the convenient approximate representation of Fig. 34 yields results that are entirely satisfactory from a quantitative point of view. It will be noted that the distribution curves at *a* and *b* of Fig. 33 would be half sine waves if the minima reached zero. As a result the curves of Fig. 34 are sine wave in character, and this fact can be used in solving problems involving

distributions in circuits with distributed constants, provided, of course, that the effect of the finite though small minima can be neglected.

Voltage and Current Distribution with Any Type of Load Impedance.—When the receiving end of a circuit with distributed constants is closed through a resistance of $\sqrt{L/C}$ ohms, where L and C are, respectively, the inductance and capacity per unit length of circuit, the voltage and current distribution is as shown in Fig. 33*c* and the impedance at the sending end is a constant resistance of $\sqrt{L/C}$ ohms for all frequencies. This particular value of receiver resistance is called the *characteristic impedance* of the line and destroys all resonances, no matter what line length or frequency is being considered.

When the load at the receiving end of the line is a resistance that is greater than the characteristic impedance, but is not infinite, the type of voltage and current distribution that can be expected is shown in Fig. 33*d*. The resonances that now exist are similar in general character to those with open circuit at the receiver but differ in that the variation that takes place between adjacent quarter-wave-length points is not so great. If the load at the receiver is a resistance which is less than the characteristic impedance, but is not a short circuit, the distribution obtained is similar in its general character to the short-circuit case, but the resonance effects at the quarter-wave-length points are not so pronounced. The voltage and current distributions are now similar to those shown for current and voltage, respectively, in Fig. 33*d*.

With reactive loads the first resonance point comes closer to the receiver than a quarter of a wave length, as shown in Figs. 33*e* and 33*f*. With capacitive loads the distribution is essentially of the same character as with open circuit, except that the curves are all shifted bodily toward the receiver by an amount that becomes greater as the capacitive reactance is reduced. With inductive loads the distribution is essentially of the same character as with short circuit, except that everything is shifted toward the receiver by an amount that increases as the load reactance approaches an open circuit. These displacements in position with reactive loads are clearly brought out in Fig. 33.

Transmission lines that are either open- or short-circuited at the receiving end have many of the properties of ordinary resonant circuits. Thus, if the line is open at the receiver and is an odd number of quarter wave lengths long, the receiving voltage is much higher than the sending-end voltage, giving a resonant rise of voltage similar to that obtained with a series circuit. Also, if the line is an even number of quarter wave lengths long and is open at the receiver, the sending-end current is low and the line offers a high impedance at its sending end that has the same general characteristics as the impedance of an ordinary parallel-resonant circuit. When properly designed, resonant transmission lines have very high

effective Q's at high frequencies and are superior to tuned circuits consisting of an ordinary coil and condenser.[1]

16. Inductively Coupled Circuits; Theory.—When mutual inductance exists between coils that are in separate circuits, these circuits are said to be inductively coupled. The effect of the mutual inductance is to make possible the transfer of energy from one circuit to the other by transformer action. That is, an alternating current flowing in one circuit produces magnetic flux which induces a voltage in the coupled circuit, resulting in induced currents and a transfer of energy from the first or primary circuit to the coupled or secondary circuit. Several types of inductively coupled circuits commonly encountered in radio work are shown in Fig. 35.

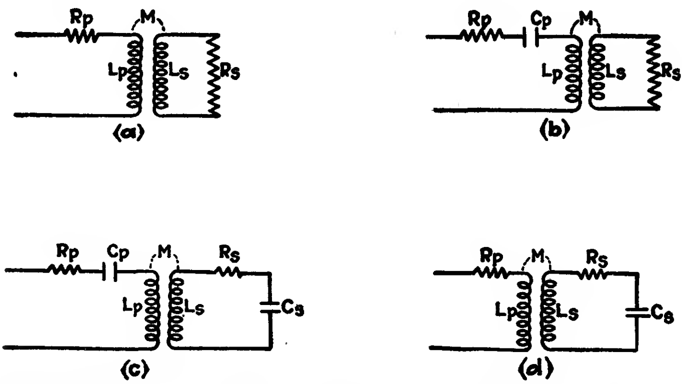


FIG. 35.—Various types of inductively coupled circuits commonly encountered in radio work.

The behavior of inductively coupled circuits is somewhat complex, but it can be readily calculated with the aid of the following rules.

Rule 1. *As far as the primary circuit is concerned, the effect that the presence of the coupled secondary circuit has is exactly as though an impedance* $(\omega M)^2/Z_s$ *had been added in series with the primary,*[2] *where*

M = mutual inductance

$\omega = 2\pi f$

Z_s = series impedance of secondary circuit when considered by itself.

[1] For a complete discussion of the properties of resonant transmission lines used to perform the functions of ordinary resonant circuits, see F. E. Terman, Resonant Lines in Radio Circuits, *Elec. Eng.*, vol. 53, p. 1046, July 1934.

[2] This can be demonstrated by writing down the circuit equations for the primary and secondary. These equations are

$$E = I_pZ_p + j\omega MI_s$$
$$\text{Induced voltage} = -j\omega MI_p = I_sZ_s$$

The equivalent impedance $(\omega M)^2/Z_s$ which the presence of the secondary adds to the primary circuit is called the *coupled impedance,* and, since Z_s is a vector quantity having both magnitude and phase, the coupled impedance is also a vector quantity, having resistance and reactance components.

Rule 2. *The voltage induced in the secondary circuit by a primary current of I_p has a magnitude of ωMI_p and lags the current that produces it by* 90°. *In complex quantity notation the induced voltage is $-j\omega MI_p$.*

Rule 3. *The secondary current is exactly the same current that would flow if the induced voltage were applied in series with the secondary and if the primary were absent.*[1] The secondary current therefore has a magnitude $\omega MI_p/Z_s$, and in complex quantity representation is given by $-j\omega MI_p/Z_s$.

These three rules hold for all frequencies and for all types of primary and secondary circuits, both tuned and untuned. The procedure to follow in computing the behavior of a coupled circuit is, *first,* to determine the primary current with the aid of Rule 1; *second,* to compute the voltage induced in the secondary, knowing the primary current and using Rule 2; *finally,* to calculate the secondary current from the induced voltage by means of Rule 3. The following set of formulas will enable these operations to be carried out systematically.

$$\left.\begin{array}{l}\text{Impedance coupled into primary}\\ \text{circuit by presence of the secondary}\end{array}\right\} = \frac{(\omega M)^2}{Z_s} \tag{38}$$

$$\text{Equivalent primary impedance} = Z_p + (\omega M)^2/Z_s \tag{39}$$

---

where Z_p is the series impedance of the primary and E is the voltage applied to the primary. Solving this pair of equations to eliminate I_s gives

$$E = I_p\left[Z_p + \frac{(\omega M)^2}{Z_s}\right] \tag{37}$$

This relation shows that the effective primary impedance with secondary present is $Z_p + \frac{(\omega M)^2}{Z_s}$, of which the second term represents the coupled impedance arising from the presence of the secondary.

[1] Some readers may wonder why it is that, although the secondary circuit couples an impedance into the primary, the primary is not considered as coupling an impedance into the secondary. The explanation for this is as follows: The effect that the secondary really has upon the primary circuit is to induce a back voltage in the primary proportional to the secondary current. This back voltage represents a voltage drop occurring in the primary circuit which is equivalent to the voltage drop produced across the hypothetical coupled impedance considered as being in series with the primary. The impedance that the secondary couples into the primary is hence a means of taking into account the voltage the secondary current induces in the primary. The voltage that is induced in the secondary circuit by the primary current is taken into account by Rule 3, so that no coupled impedance need be postulated as present in the secondary to take into account the effect of the primary.

$$\text{Primary current} = I_p = \frac{E}{Z_p + \frac{(\omega M)^2}{Z_s}} \tag{40}$$

$$\text{Voltage induced in secondary} = -j\omega M I_p \tag{41}$$

$$\text{Secondary current} = \frac{-j\omega M I_p}{Z_s} = \frac{-j\omega M E}{Z_p Z_s + (\omega M)^2} \tag{42}$$

In these equations:

M = mutual inductance between primary and secondary

$\omega = 2\pi f$

Z_s = series impedance of secondary circuit considered as though primary were removed

Z_p = series impedance of primary circuit considered as though secondary were removed

E = applied voltage.

The primary and secondary impedances Z_p and Z_s, respectively, are vector quantities, so that Eqs. (38) to (42) are consequently all vector equations. Thus in the case of the coupled circuits of Fig. 35*c* one has:

$$Z_p = R_p + j\left(\omega L_p - \frac{1}{\omega C_p}\right)$$

$$Z_s = R_s + j\left(\omega L_s - \frac{1}{\omega C_s}\right)$$

For other circuits the impedances Z_p and Z_s are given by expressions similar to these, but differing in detail.

Action of the Coupled Impedance.—Many of the important properties of coupled circuits can be determined by examining the nature of the coupled impedance $(\omega M)^2/Z_s$. When the mutual inductance M is very small, the impedance coupled into the primary by the presence of the secondary is correspondingly small, and the primary current is very nearly the same as though no secondary were present. The coupled impedance is also very small when the secondary impedance Z_s is large because, even when a large mutual inductance induces a high voltage in the secondary, the secondary current produced is small as a result of the high impedance, and little energy transfer takes place. When the secondary impedance Z_s is low and the mutual inductance is not too small, the coupled impedance $(\omega M)^2/Z_s$ is large and the voltage and current relations in the primary circuit are affected to a considerable extent by the presence of the coupled secondary. In particular, when the secondary is a series-resonant circuit, as in Figs. 35*c* and 35*d*, the secondary impedance Z_s will be very low at the resonant frequency of the secondary, so that at this frequency the coupled impedance will be very high, and the presence of the secondary will be an important factor in determining the primary current.

The coupled impedance has the same phase angle as the secondary impedance Z_s, with the exception that the sign of the angle is reversed; that is, if the secondary impedance is inductive and has an angle of 30° lagging, the impedance coupled in series with the primary circuit by the action of the secondary has a phase angle of 30° leading. The physical significance of this change from lagging to leading is that coupling a secondary circuit having an inductive reactance to a primary circuit is equivalent to neutralizing some of the inductive reactance already possessed by the primary, and this is done electrically by postulating a capacitive reactance of suitable magnitude in series with the inductance to be neutralized.[1] When the secondary impedance Z_s is a pure resistance, the coupled impedance will also be a resistance. This is a very important special case because, when the secondary is a tuned circuit, as in Figs. 35*c* and 35*d*, the secondary impedance will be a resistance at resonance.

The energy consumed by the secondary circuit is the energy represented by the primary current flowing through the resistance component of the coupled impedance. In the same way the reactive volt-amperes transferred from the primary to the secondary are the volt-amperes developed by the primary current flowing through the reactive part of the coupled impedance. By expressing the secondary impedance Z_s in terms of its real and reactive components R_s and X_s, respectively, one can rewrite Eq. (38) in the following form:

$$\text{Coupled impedance} = \frac{(\omega M)^2 R_s}{R_s^2 + X_s^2} - j\frac{(\omega M)^2 X_s}{R_s^2 + X_s^2} \tag{43}$$

in which the resistance component of the coupled impedance is $\dfrac{(\omega M)^2 R_s}{R_s^2 + X_s^2}$ and the reactance component is $-j\dfrac{(\omega M)^2 X_s}{R_s^2 + X_s^2}$. The effect that the secondary has on the primary circuit is then exactly as though this resistance and this reactance had been inserted in series with the primary circuit and as though this resistance and reactance had consumed the energy and the reactive volt-amperes transferred to the secondary.

It will be noted that, although two inductance coils between which mutual inductance exists constitute a transformer, the inductively coupled circuit is not treated here in terms of leakage reactances, turn ratio, magnetizing current, and so on, as are power transformers. The reason for this is that in transformers used for radio work the leakage reactances are very high, causing the ratio of voltage transformation to be entirely different from the turn ratio.

[1] Although the coupled impedance is capacitive and so neutralizes part of the primary inductance, it is impossible to obtain a resultant capacitive reactance in the primary circuit by very large coupling since, with the maximum coupling that can possibly exist ($k = 1$), it will be found that the coupled capacitive reactance will just neutralize the inductive reactance of the primary and no more.

The conventional equivalent circuit of a transformer in terms of the inductance and coefficient of coupling is shown in Fig. 36. Here the total primary inductance is broken up into a leakage inductance L' and a coupled inductance L_c', while the secondary is likewise broken up into leakage inductance L'' and a coupled inductance L_c''. The leakage inductance is considered as having no coupling whatsoever to the other winding, while the coupled inductances L_c' and L_c'' are taken as having a coefficient of coupling equal to unity. The values of these inductance components in terms of the coefficient of coupling and the primary, secondary, and mutual inductances are given in the figure. It will be noted that, when the coefficient of coupling is low, for example, 0.01, then practically all the primary and secondary inductances are leakage

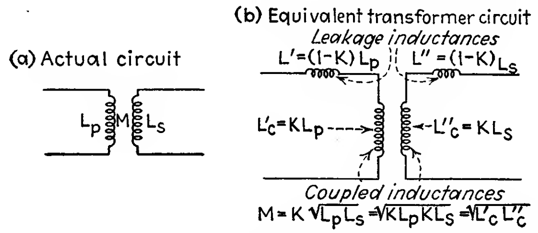


FIG. 36.—Equivalent circuit of a transformer expressed in terms of primary and secondary inductances, mutual inductance, and coefficient of coupling.

inductances and the power-transformer method of representation is not particularly convenient.

17. Analysis of Some Typical Coupled Circuits.—In this section the types of coupled circuits most commonly encountered in communication work will be analyzed by the principles given above.

Coupled Circuit with an Untuned Secondary Consisting of a Resistance and Inductance.—This arrangement is illustrated by Fig. 35*a* and is the type of coupled circuit encountered when the secondary is a mass of metal such as a shield can, a metal panel, screw, mounting bracket, etc., which is near the coil representing the primary circuit. Such a secondary is essentially an inductance in series with a resistance and so has a lagging reactance. As a consequence the secondary couples into the primary a resistance component and also a capacitive component. The effect of the coupled resistance is to increase the effective resistance that appears between the primary terminals. The effect of the coupled capacitive reactance is to neutralize a portion of the primary inductance, thereby reducing the equivalent inductance that is observed between the terminals of the primary coil. With perfect coupling ($k = 1.0$), the primary inductance would be completely neutralized, while with lesser degrees of coupling the effect of the secondary upon the primary inductance is correspondingly less.

When the impedance of the secondary is largely reactive, that is to say, when the secondary is a metal object of low-resistance material, the resistance component of the coupled impedance is small and the principal effect upon the primary coil is the reduction in effective inductance which is produced by the coupled reactance. Consequently, if a coil must be located near a metal object such as a shield can or a panel, the metal used should be the best possible conductor, preferably copper or aluminum, since in this way the added losses will be small. If the resistance of the secondary can be assumed to be zero, the coupled impedance is a capacitive reactance having no resistance component, and the equivalent primary inductance is

$$\left.\begin{array}{l}\text{Equivalent primary inductance}\\ \text{in presence of secondary}\end{array}\right\} = L_p(1 - k^2) \tag{44}$$

where L_p is the primary inductance with the secondary removed and k is the coefficient of coupling between the coil and the secondary. In order to calculate the reduction in effective inductance which is produced by a shield or a metal panel, one needs to know only the coefficient of coupling. The number of primary turns, etc., is not important.

Coupled Circuits with Untuned Primary and Tuned Secondary.—A circuit of this type is shown in Figs. 35*d* and 37 and is the fundamental circuit of the transformer-coupled tuned radio-frequency amplifier.

As the circuit is ordinarily encountered, the primary resistance R_p represents the plate resistance of a tube and is much larger than the reactance ωL_p of the primary inductance. A curve showing the variation of secondary current with frequency in a typical case is shown by the solid lines in Fig. 37 for the case where the primary resistance is much higher than the primary reactance. The resulting curve is seen to have the same shape as an ordinary resonance curve with the maximum at the resonant frequency of the secondary, but upon close inspection it will be found to correspond to a somewhat lower Q than that of the secondary circuit.

When the reactance ωL_p of the primary inductance is not negligible compared with the primary resistance, the curve of secondary current as a function of frequency still has the shape of a resonance curve, but the frequency at which the secondary current is maximum has been shifted to a frequency higher than the resonant frequency of the secondary. This is shown by the dotted lines in Fig. 37 for a typical case.

A quantitative analysis of the circuit arrangement shown in Fig. 37 can be obtained by the use of Eqs. (38) to (42). The coupled impedance is given by the expression

$$\text{Coupled impedance} = \frac{(\omega M)^2}{Z_s} = \frac{(\omega M)^2}{R_s + j\left(\omega L_s - \dfrac{1}{\omega C_s}\right)}$$

An examination of this expression shows that in the limited frequency range in which the resonance effects take place the numerator is substantially constant, whereas the denominator represents the series impedance of the secondary circuit. This is therefore an equation of the same general type as Eq. (30*b*) for parallel resonance. *The coupled impedance produced by a tuned secondary circuit consequently varies with frequency according to the same general law as does the parallel impedance of the secondary circuit.* The absolute magnitude of the curve, however, depends upon the mutual inductance.

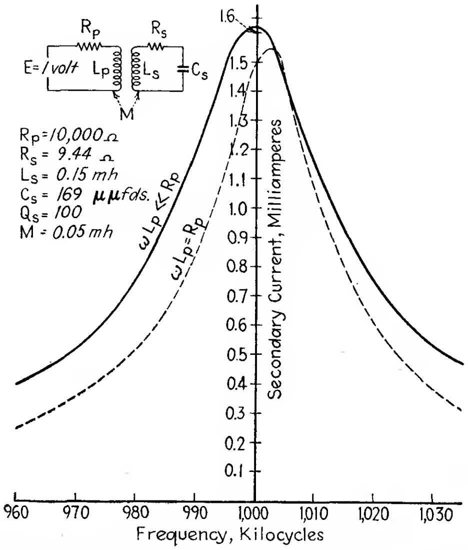


FIG. 37.—Secondary current when the secondary is a resonant circuit and the primary is untuned. The solid lines are for the case where the primary inductance has negligible reactance, while the dotted lines are for a large primary inductance but the same mutual inductance. It will be observed that in both cases the secondary current follows a resonance curve, but the shape corresponds to a lower Q than that of the secondary circuit taken alone.

Two Resonant Circuits Tuned to the Same Frequency and Coupled Together.—When two resonant circuits that are tuned to the same frequency are coupled together, the resulting behavior depends very largely upon the degree of coupling, as seen from Fig. 38. When the coefficient of coupling is small, the curve of primary current as a function of frequency is substantially the series-resonance curve of the primary circuit considered alone, while the secondary current is small and varies with frequency in such a way as to be much more peaked than the resonance curve of the secondary circuit considered as an isolated circuit. As the coefficient of coupling is increased somewhat, the curve of primary current becomes broader, as a result of a reduction in the primary current at resonance and an increase in the primary current at frequencies slightly

off resonance. At the same time the secondary current peak becomes larger and the curve of secondary current somewhat broader. These trends continue as the coefficient of coupling is increased until the coupling is such that the resistance which the secondary circuit couples into the primary is equal to the primary resistance. This is called the *critical coupling*, and it causes the secondary current to have the maximum value

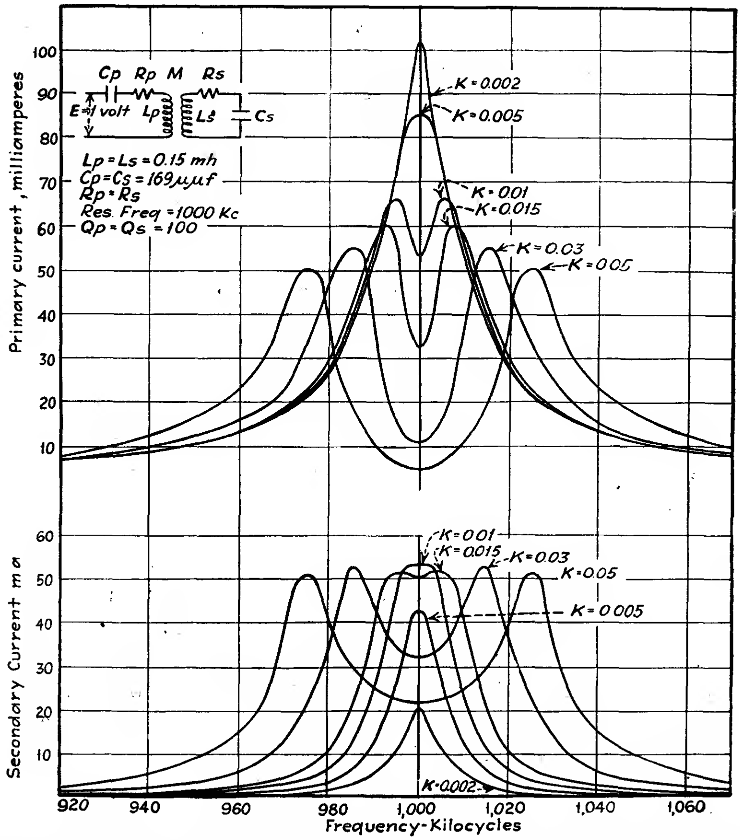


FIG. 38.—Curves for two circuits separately tuned to the same frequency and coupled together, showing variation of primary and secondary currents with frequency for a number of coefficients of coupling.

it can attain. The curve of secondary current is then somewhat broader than that of the resonance curve of the secondary circuit considered above. The primary current now has two peaks since the primary current is slightly more at frequencies just off resonance than at the resonant frequency. As the coefficient of coupling is increased beyond the critical value, the double humps in the primary current become more prominent and the peaks spread farther apart. The curve of secondary current also

begins to display double humps, with the peaks becoming more pronounced and spreading farther apart as the coupling increases. The value of the primary current at the peaks becomes smaller the greater the coupling, but in the secondary circuit the peaks have substantially the same value irrespective of the coefficient of coupling provided the coupling is not less than the critical value.

The reason for the above behavior follows from the theory of coupled circuits and centers around the way in which the coupled impedance $(\omega M)^2/Z_s$ varies with frequency. Consider first the total primary circuit impedance. This consists of the actual self-impedance of the primary plus whatever impedance the secondary circuit couples into the primary. The type of coupled impedance produced by a tuned secondary has already been discussed; it is substantially a parallel-resonance curve having a shape corresponding to the Q of the secondary circuit and an amplitude determined by the mutual inductance. The coupled impedance is hence maximum at resonance and is then a resistance. At frequencies below resonance the coupled impedance is inductive and at frequencies above resonance it is capacitive, as shown in Fig. 32.

When this coupled impedance is added to the self-impedance of the primary circuit, the effect at resonance is to increase the effective primary resistance above the value that would exist in the absence of the secondary. At frequencies somewhat below resonance the coupled impedance is largely inductive whereas the primary self-impedance is capacitive, so that the coupled inductive reactance neutralizes some of the primary capacitive reactance, lowering the primary circuit impedance and increasing the primary current. The situation is somewhat similar for frequencies above resonance except that now the coupled reactance is capacitive and neutralizes some of the inductive reactance which the primary circuit otherwise has at frequencies above resonance. Consequently the net effect of the coupled impedance is to lower the primary current at the resonant frequency and raise the current at frequencies somewhat off resonance. The magnitude of this effect depends upon the coefficient of coupling, being naturally small when the coupling is loose. When the coupling is of the order of magnitude of the critical value or greater, the coupled impedance becomes sufficient to be the major factor in determining the impedance of the primary circuit. In particular, at resonance the primary current tends to be small because of the very large coupled resistance, while there is a frequency on each side of resonance at which the coupled reactance exactly neutralizes the primary reactance, giving zero reactance for the total primary circuit impedance and causing the flow of a large current. This is the cause of the double-humped curves of primary current for high couplings, such as shown in Fig. 38.

The curve of secondary current is determined by the secondary impedance, and by the voltage induced in the secondary by the primary current. The induced voltage varies with frequency in almost exactly the same way as does the primary current, since the induced voltage is equal to ωMI_p; and in the limited frequency range in which the resonance effects take place, ω changes very little. As a result of this the curve of secondary current has a shape that is almost exactly the product of the shape of the curve of primary current and the shape of the resonance curve of the secondary circuit. Since the latter curve is sharply peaked, the secondary current is much more peaked than the primary current, as is clearly evident in Fig. 38. At low coefficients of coupling the curve of secondary current is particularly sharp, being substantially the product of the resonance curves of the primary and secondary circuits. As the coupling increases, the primary-current curve becomes broader, thereby making the secondary curve less sharp. At the same time the peak amplitude of the secondary current increases because of the increased coupling. When the coefficient of coupling reaches the critical value, the secondary current has the maximum value it can attain, and the curve of secondary current in the immediate vicinity of resonance is somewhat broader on top than the secondary resonance curve because of the double-humped curve of primary current that exists with critical coupling. As the coupling is increased beyond the critical value, the secondary-current peak splits into two peaks which have amplitudes substantially the same as the secondary-current peak at critical coupling. The separation between these peaks increases with frequency, and is substantially the same as the separation of the peaks of primary current.

The exact shape of the curves shown in Fig. 38 can be calculated with the aid of Eqs. (38) to (42). In practical work the most important quantity is the secondary current, and it is often desirable to rearrange Eq. (42) to express the secondary current in terms of the circuit Q's, the coefficient of coupling, and the ratio of the actual frequency to the resonant frequency. When this is done, one obtains[1]

[1] The rearrangement of Eq. (42) to give Eq. (45) is accomplished by starting with the final form of Eq. (42) and making the following substitutions:

$$Z_p = R_p + j\left(\omega L_p - \frac{1}{\omega C_p}\right) = R_p + j\omega L_p\left(1 - \frac{1}{\omega^2 L_p C_p}\right) = R_p + j\omega L_p\left(1 - \frac{1}{\gamma^2}\right)$$

$$Z_s = R_s + j\left(\omega L_s - \frac{1}{\omega C_s}\right) = R_s + j\omega L_s\left(1 - \frac{1}{\omega^2 L_s C_s}\right) = R_s + j\omega L_s\left(1 - \frac{1}{\gamma^2}\right)$$

This gives

$$I_s = \frac{-jE\omega M}{R_pR_s - \left(1 - \frac{1}{\gamma^2}\right)^2 \omega^2 L_p L_s + (\omega M)^2 + j\left(1 - \frac{1}{\gamma^2}\right)(\omega L_p R_s + \omega L_s R_p)}$$

Dividing both numerator and denominator by $\omega^2 L_p L_s$, noting that $Q_p = \omega L_p/R_p$, $Q_s = \omega L_s/R_s$, $M^2/L_pL_s = k^2$, and $\omega = \gamma\omega_0$, results in Eq. (45).

Secondary current $= I_s =$

$$\frac{-jEk}{\gamma\omega_0\sqrt{L_pL_s}\left[k^2 + \frac{1}{Q_pQ_s} - \left(1 - \frac{1}{\gamma^2}\right)^2 + j\left(1 - \frac{1}{\gamma^2}\right)\left(\frac{1}{Q_p} + \frac{1}{Q_s}\right)\right]} \tag{45}$$

where

$\omega_0 =$ 2π times resonant frequency

$\gamma = \dfrac{\text{actual frequency}}{\text{resonant frequency}}$

$E =$ voltage applied in series with primary

$Q_p = \omega L_p/R_p$ for primary circuit

$Q_s = \omega L_s/R_s$ for secondary circuit

$k =$ coefficient of coupling

$L_p =$ total inductance of primary circuit

$L_s =$ total inductance of secondary circuit.

Since this equation is too complicated to permit qualitative analysis by inspection, it is convenient to derive special formulas giving the most important features involved in the curves of secondary current. The points that are of greatest interest are the conditions required for critical coupling, the secondary current at resonance, the secondary current at the peaks of the secondary current curves if the coupling is more than the critical value, and the separation of these peaks of secondary current.

The secondary current reaches its maximum possible value when the circuits are in resonance and when the coupled impedance is a resistance equal to the resistance of the primary circuit. This condition exists when $(\omega M)^2/R_s = R_p$, or

$$(\omega M) = \sqrt{R_pR_s} \tag{46}$$

The coefficient of coupling k required to give the critical coupling can be expressed in terms of the ratios $Q_p = \omega L_p/R_p$ and $Q_s = \omega L_s/R_s$ of the primary and secondary circuits, respectively, by using these two relations to eliminate R_p and R_s which appear in Eq. (46). When this is done, the critical coefficient of coupling is found to be

$$\text{Critical } k = \frac{1}{\sqrt{Q_pQ_s}} \tag{47}$$

The critical coefficient of coupling is usually very small since, if the circuit Q's are 100, the corresponding coefficient is only 0.01.

The secondary current at resonance is the current given by Eq. (45) when γ equals unity:

$$\text{Secondary current at resonance} = -j\frac{E_0k}{\omega_0\sqrt{L_pL_s}\left(k^2 + \frac{1}{Q_pQ_s}\right)} \tag{48a}$$

When the coefficient of coupling has the critical value, the secondary current is at a maximum and can be found by substituting Eq. (47) into Eq. (48a), giving

$$\left.\begin{array}{l}\text{Secondary current at resonance}\\ \text{with critical coupling}\end{array}\right\} = -j\frac{E_0\sqrt{Q_pQ_s}}{2\omega_0\sqrt{L_pL_s}} = -j\frac{E_0}{2\sqrt{R_pR_s}} \tag{48b}$$

When the coupling is greater than the critical value, the secondary-current curve has two humps. The current at these peaks is the same as the resonant current with critical coupling, as has already been mentioned, and is given by Eq. (48b). The spacing of the humps depends primarily upon the coefficient of coupling, but it is also influenced somewhat by the ratio of this actual coefficient of coupling to the critical coupling. When the coupling is considerably greater than the critical value, the peaks occur at practically the same frequencies as if the circuit resistances were zero. One then has[1]

$$\frac{\text{Frequency at secondary peak}}{\text{Resonant frequency of tuned circuits}} = \frac{1}{\sqrt{1 \pm k}} \tag{49a}$$

The plus sign denotes the lower frequency and the minus sign the higher frequency peak.

The voltage developed across the secondary condenser is equal to the reactance of this condenser times the secondary current; thus it can readily be calculated once the current curve is known. For most purposes it is

[1] Equation (49a) is derived by starting with Eq. (45) and assuming Q_p and Q_s are infinite ($i.e.$ assuming zero losses). This gives

$$\text{Secondary current} = \frac{-jE_0k}{\gamma\omega_0\sqrt{L_pL_s}\left[k^2 - \left(1 - \frac{1}{\gamma^2}\right)^2\right]}$$

The secondary current becomes infinite (assuming zero losses) when the denominator is zero, or when

$$k = \left(1 - \frac{1}{\gamma^2}\right)$$

Solving this for γ gives Eq. (49a), since the value of γ obtained is the ratio of frequency required to give maximum secondary current to the resonant frequency.

If the effect of the resistance upon the location of the peaks is taken into account, a complicated manipulation based on Eq. (45) gives

$$\frac{\text{Frequency at secondary peak}}{\text{Resonant frequency of tuned circuits}} = \frac{1}{\sqrt{1 \pm k\left[1 - \frac{1}{2k^2}\left(\frac{1}{Q_p^2} + \frac{1}{Q_s^2}\right)\right]^{1/2}}} \tag{49b}$$

Equation (49b) is the general equation while Eq. (49a) is a special form for the case where Q_p and Q_s are infinite ($i.e.$, for zero losses). The latter equation is commonly used for all cases, however, because of its much greater simplicity and the fact that it still gives a general idea of the width between the peaks of secondary current.

sufficient to assume that the curve of voltage developed across the condenser has the same shape as the curve of secondary current. One is interested primarily in the behavior about resonance, and in the limited frequency range consequently involved the condenser reactance changes very little.

The calculation of complete curves of secondary current, such as given by Fig. 38, is rather involved and usually not worth the trouble. For many practical purposes it is sufficient to calculate only a few critical characteristics. The first of these are the actual coefficient of coupling and the critical coefficient of coupling, the latter obtainable by Eq. (47). A comparison of these two will give the type of behavior to be expected. One can then calculate the response at resonance, using Eq. (48*a*); if the coupling exceeds the critical value, one can determine the location of the peaks of secondary current by Eq. (49*a*) or (49*b*) and the magnitudes of these peaks by Eq. (48*b*).

Two Resonant Circuits Tuned to Slightly Different Frequencies and Coupled Together.—This differs from the case just discussed in that the primary and secondary circuits are adjusted to have slightly different resonant frequencies, *i.e.*, are detuned. The behavior then depends upon both the amount of detuning and the degree of coupling. Curves showing the variation of primary and secondary currents as a function of frequency for some typical cases are shown in Fig. 39. Examination of these shows that the secondary current curves are symmetrical about a frequency midway between the individual resonant frequencies, and that the curves have the same general shape as though the circuits were tuned to the same frequency and had a somewhat higher coupling than used in Fig. 39. Quantitative analysis confirms this, and shows that to a good approximation the shape of the secondary current curve that is obtained with detuning is the same as the shape that would be obtained if the circuits were both tuned to the same frequency and the coefficient of coupling were increased to $\sqrt{k^2 + \left(\frac{\Delta}{f_0}\right)^2}$, where k is the actual coefficient of coupling, Δ is the difference between the resonant frequencies of the primary and secondary circuits, and f_0 is the frequency midway between the primary and secondary resonant frequencies.[1] *Hence detuning primary and secondary circuits slightly has approximately the same effect on the shape of the secondary current curve as increasing the coefficient of coupling when there is no detuning.* The only essential difference between the curves of secondary current obtained by detuning and the

[1] See Harold A. Wheeler and J. Kelly Johnson, High Fidelity Receivers with Expanding Selectors, *Proc. I.R.E.*, vol. 23, p. 594, June, 1935. The principal assumptions involved are that Δ/f_0 is small compared with unity and that the primary and secondary circuits have the same value of Q.

curves obtained by use of increased coupling is that in the latter case the absolute magnitudes may be greater.

The primary current obtained when the primary and secondary circuits are detuned is unsymmetrical, as shown in Fig. 39. There is always a peak at a frequency that is very nearly the resonant frequency of the primary circuit, and in addition one can expect a certain amount of deformation in the curve near the frequency at which the secondary is resonant. This latter effect occurs as a result of the impedance that the secondary

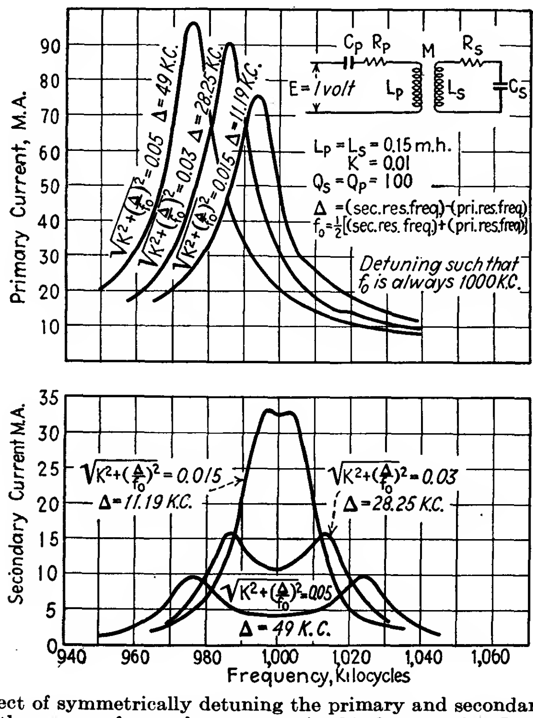


FIG. 39.—Effect of symmetrically detuning the primary and secondary circuits. Note the similarity in the curves of secondary current in this figure and in Fig. 38, showing how detuning has the same effect on the shape of the secondary current curve as increasing the coefficient of coupling.

couples into the primary, which is maximum at the secondary resonant frequency but small as in Fig. 39 unless the mutual inductance is large.

18. Miscellaneous Coupling Methods.—Energy can be transferred from one circuit to another by a variety of coupling methods, in addition to the inductive coupling just considered. Thus in Fig. 10*b* the coupling consists of an inductance common to two circuits while in Fig. 10*c* the coupling is provided by a capacity common to two circuits. Still other coupling methods of a relatively simple nature are illustrated in Fig. 40, while complex coupling networks are given in Fig. 124. The behavior of

all these coupled circuits follows the same general character as that discussed for inductive coupling. Thus the secondary circuit can be considered as producing an equivalent coupled impedance in the primary circuit while the primary circuit can be considered as inducing in the secondary a voltage that gives rise to the secondary current. The analysis may become relatively complicated, however, particularly in the case of complex coupling networks such as shown in Fig. 124.

The simplest method of analyzing these various forms of coupled circuits is to take advantage of the fact that all of them can be reduced to the simple coupled circuit of Fig. 40*g*, provided suitable values are assigned

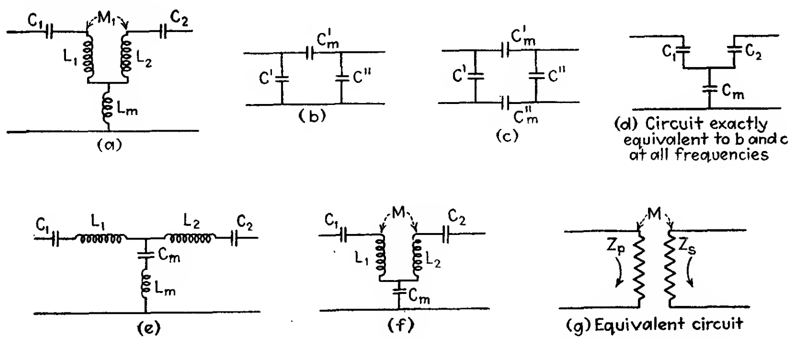


FIG. 40.—Some of the methods that may be used in coupling circuits.

to Z_p, Z_s, and M. The rules that determine the values of these quantities in the simple equivalent circuit are as follows:

1. The equivalent primary impedance Z_p of the equivalent circuit is the impedance that is measured across the primary terminals of the actual circuit when the secondary circuit has been opened.
2. The secondary impedance Z_s of the equivalent circuit is the impedance that is measured by opening the secondary of the actual circuit and determining the impedance between these open points when the primary is open-circuited.
3. The equivalent mutual inductance M is determined by assuming a current I_0 flowing into the primary circuit. The voltage which then appears across an open circuit in the secondary is equal to $-j\omega MI_0$.

In making use of the equivalent circuit of Fig. 40*g* it is to be remembered that the values of Z_p, Z_s, and M may all vary with frequency, so that it is generally necessary to determine a new equivalent circuit for each frequency at which calculations are to be made.

After the actual coupled circuit has been reduced by the above procedure to its equivalent form shown in Fig. 40*g*, one can then apply the formulas that have already been derived for inductively coupled circuits, using the appropriate values M, Z_s, Z_p as determined for the equivalent circuit. This procedure has the advantage of using the same fundamental formulas to handle all types of coupling, and makes it possi-

ble to carry on the analysis in the same manner for all cases. The method is particularly convenient in the handling of complex coupling networks such as illustrated in Fig. 124.

The quantity M that appears in the equivalent circuit represents the effective coupling that is present between the primary and secondary circuits. It is not necessarily a real mutual inductance of the inductive type, but rather a sort of mathematical fiction that gives the equivalent effect of whatever coupling is really present. If the actual coupling is capacitive, the numerical value of M will be found to be negative; if the coupling is of a complex type representing both resistive and reactive coupling, the numerical value of M will be found to have both real and imaginary parts. This need introduce no uncertainty, however, since the proper procedure is to take the value of M as it comes and substitute it with its appropriate sign and phase angle whenever M appears in the expressions previously derived for inductively coupled circuits.

When this analysis is applied to the capacitively coupled circuits of Figs. 10*c*, 40*b*, 40*c*, and 42*c* the results are essentially the same as for inductive coupling, since the coefficient of coupling is independent of frequency. Thus, when primary and secondary are both tuned to the same frequency in Fig. 42*c*, the secondary-current characteristic has two humps if the coupling is large, *i.e.*, if condenser C_m is small, while there is only one peak of secondary current when the coupling is small, *i.e.*, when condenser C_m is large.

Circuits having combined electromagnetic and electrostatic coupling, such as those at *e* and *f* of Fig. 40, behave as ordinary coupled circuits except that the coefficient of coupling varies with frequency. Thus in the case of circuit *e* the circuit is capacitively coupled at low frequencies and inductively coupled at high frequencies because the coupling combination of C_m in series with L_m has capacitive and inductive reactance under these respective conditions. In between, at the resonant frequency of L_m and C_m, there is no coupling and $k = 0$. The arrangement shown at *f* acts similarly as a circuit with a coefficient of coupling that varies with frequency. Depending upon the relative direction in which primary and secondary are wound in *f*, the mutual inductance M may introduce an inductive coupling that either aids or opposes the capacitive coupling. Circuits having combined electrostatic and electromagnetic coupling find application where it is desired to obtain a coefficient of coupling that varies with frequency, as is commonly the case in tuned amplifiers and antenna-coupling circuits of radio receivers.[1]

[1] For further information see: Harold A. Wheeler and W. A. MacDonald, Theory and Operation of Tuned Radio-frequency Coupling Systems, *Proc. I.R.E.*, vol. 19, p. 738, May, 1931; Edwin H. Loftin and S. Young White, Combined Electromagnetic and Electrostatic Coupling, and Some Uses of the Combination, *Proc. I.R.E.*, vol. 14, p. 605, October, 1926.

19. Band-pass Filters.[1]—Examination of Fig. 38 shows that it is possible, by suitably coupling together two circuits tuned to the same frequency, to obtain a curve of secondary current which is substantially constant over a limited range of frequencies around resonance and which then falls off rapidly at frequencies farther off resonance. When this result is obtained, one is said to have a *band-pass filter*. Such band-pass characteristics are particularly desirable in handling modulated waves because the response is practically the same to all side-band frequencies contained in the wave as to the carrier. In contrast with this, ordinary resonant circuits have a rounded top and so discriminate against the higher side-band frequencies in favor of the lower side-band frequencies and the carrier.

The most important characteristics of a band-pass filter are the width of the band of frequencies which is transmitted and the uniformity of the response within this band. When the band-pass filter is properly designed the width of the pass band is approximately the distance between the coupling humps when the circuit resistance is low. By making use of Eq. (49*a*) one gets

$$\frac{\text{Width of pass band}}{\text{Resonant frequency of tuned circuits}} = \frac{\sqrt{1+k} - \sqrt{1-k}}{\sqrt{1-k^2}} \qquad (50a)$$

Under all ordinary conditions k is small so that this can be reduced to the approximate relation

$$\frac{\text{Width of pass band}}{\text{Resonant frequency of tuned circuits}} = k \qquad (50b)$$

The width of the pass band given by Eq. (50*b*) is slightly greater than the actual separation of the peaks of secondary current when the circuit has losses, but it represents relatively accurately the extreme limits over which the transmission is approximately uniform.

The uniformity of response within the pass band of frequencies depends on the circuit Q in relation to the coefficient of coupling. If the Q is high, the pass-band will have two peaks with a pronounced dip in the center; if the Q is low, the top will be rounded off and will look much like a broad resonance curve. For best results the circuit Q should be about 50 per cent greater than the Q that would give critical coupling with the value of k required to give the desired band width.

[1] This section is limited to a consideration of the types of band-pass filters encountered in radio work. These filters are very special cases of much more general filters which are used extensively in telephone work. For further information on telephone wave filters see F. E. Terman, "Measurements in Radio Engineering," Chap. IV, McGraw-Hill Book Company, Inc.; T. F. Shea, "Transmission Networks and Wave Filters," D. Van Nostrand Company, Inc.; E. A. Guillemin, "Communication Networks," vol. II; O. J. Zobel, Theory and Design of Uniform and Composite Wave Filters, *Bell System Tech. Jour.*, vol. 2, p. 1, January, 1923.

In designing a band-pass filter the procedure is to select first the coefficient of coupling to give the desired band width, making use of Eq. (50*b*). The circuit Q's are then made about 50 per cent greater than the Q required for critical coupling. Hence

$$\sqrt{Q_pQ_s} = \frac{1.5}{k} \tag{51}$$

It will be noted that the proper Q's are inversely proportional to the band width.

In most band-pass filter applications the useful output is the voltage developed across the condenser in the secondary circuit. The ratio of this voltage to the voltage acting in series with the primary circuit can be thought of as the resonance rise of voltage in the circuit, and it depends upon the coefficient of coupling and the Q's. When the primary and secondary circuits have the same Q, this resonance rise is equal to $Q/2$ times. Inasmuch as the resonant rise depends upon the circuit Q's, and since these must necessarily be made inversely proportional to the width of the pass band, it is seen that the resonant rise of voltage that can be realized from a band-pass filter becomes less the wider the pass band, and is only about half the resonant rise of voltage that would be obtained from one of the circuits taken alone as a simple series resonant circuit. This is the price that must be paid to secure a resonance curve having a substantially flat top combined with sides somewhat steeper than those obtained by using series resonance with a single circuit.

The effect that the circuit Q has on the uniformity of response within the pass band is brought out by Fig. 41*a*. The reduction in response caused by widening the pass band while maintaining Q at the proper value is similarly shown in Fig. 41*b*.

The results discussed in connection with Fig. 41 and Eqs. (50) and (51) are exactly the same irrespective of the method of coupling employed to give the coefficient k. The most commonly used methods are the inductive-, direct-, and capacitive-coupled arrangements shown in Fig. 42, and inductively coupled circuits which also possess capacitive coupling as a result of stray capacities. The only differences in the performances of these circuits arise from variations in the coefficient of coupling as the common resonant frequency of the two tuned circuits is changed to alter the location of the pass band.

Where a band-pass filter is used for tuning purposes instead of the usual simple series-resonant circuit, the position of the pass band is ordinarily controlled by the use of identical primary and secondary variable condensers, which are either geared together or mounted on a common shaft, so that the capacities of the two circuits remain identical as the common resonant frequency is changed. It would be possible to adjust the loca-

tion of the pass band by varying the circuit inductances, but this is seldom done because of the difficulties involved in obtaining variable inductances having low radio-frequency resistance. When tuning is accomplished by varying the circuit capacity, the coefficient of coupling is independent of

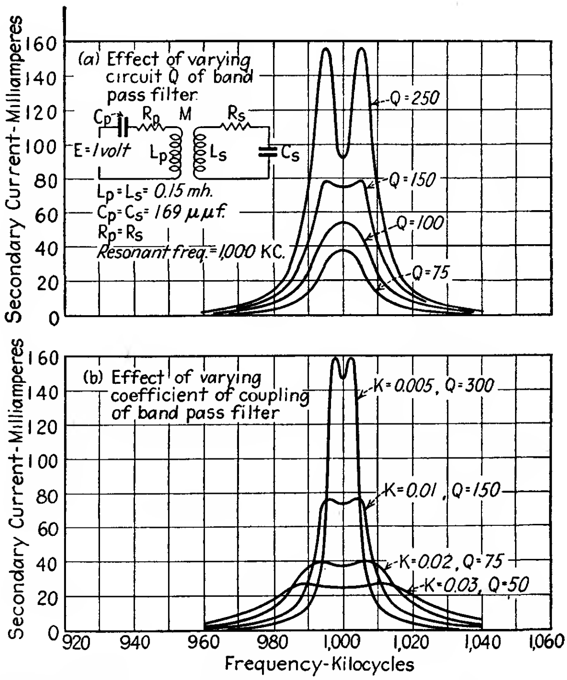


FIG. 41.—Characteristics of band-pass filter, showing (*a*) effect of circuit Q on uniformity of response within the pass band, and (*b*) effect of coefficient of coupling upon the width of the pass band and the response within the pass band when the proper circuit Q is used.

the frequency to which the circuits are tuned if inductive or direct coupling is used, but is inversely proportional to the square of the frequency with capacitive coupling.[1] Band-pass filters that are tuned by variable con-

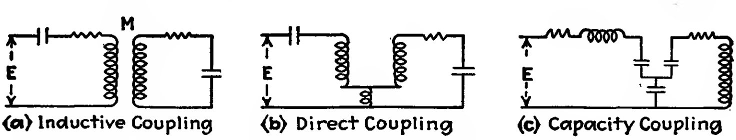


FIG. 42.—Various types of coupling employed in band-pass circuits.

densers will therefore have a width of pass band that is approximately proportional to the resonant frequency of the circuits when inductive

[1] This is readily checked by observing that the circuit capacity varies inversely as the square of the resonant frequency, so that with a constant capacity of coupling condenser the definition of coefficient of coupling leads to the result stated.

or direct coupling is used and a band width approximately inversely proportional to the resonant frequency when capacitive coupling is employed. These conclusions follow directly from Eq. (50*b*), which shows that the width of the pass band is nearly proportional to the frequency at the center of the band and also to the coefficient of coupling.

If the width of the pass band is to be constant irrespective of the frequency at which the band is located, it is necessary for the coefficient of coupling to be inversely proportional to the resonant frequency. Such a coefficient can be obtained by varying the coupling with the same control that adjusts the primary and secondary circuits, or can be approximated by a combined inductive and capacitive coupling, such as shown at *e* and *f* in Fig. 40. The effective Q required to give uniform response to the different frequencies in the pass band is seen from Eq. (51) to be inversely proportional to the coefficient of coupling. When the pass band has a constant width, the required Q will hence be proportional to the resonant frequency, which results in a resonant rise of voltage that is also inversely proportional to frequency. It is, accordingly, impossible to obtain a constant-width flat-topped pass band that has a resonant rise of voltage independent of the frequency at which the band is located. This is true irrespective of the type of coupling used or the method employed to change the location of the pass band.

Band-pass effects can also be obtained when the primary and secondary circuits are detuned slightly, as shown by the curve for $\sqrt{k^2 + (\Delta/f_0)^2} = 0.015$ in Fig. 39. The shapes of band-pass curves obtained in this way are substantially identical with those obtained in the manner described above, provided the amount of detuning and the coefficient of coupling are so related that the quantity $\sqrt{k^2 + (\Delta/f_0)^2}$ is equal to the value of k called for by Eq. (50*b*). The circuit Q's required are then the same as those called for by Eq. (51), and the general behavior is the same as though the band-pass effect was developed without detuning. The only essential difference is that detuning causes the absolute magnitude of the secondary current curve to be reduced. This makes it preferable to obtain band-pass action without resorting to detuning.

Another method of obtaining band-pass action is to combine a double-humped resonance curve with an ordinary series-resonance circuit. If the circuit Q's of a band-pass filter are much higher than the values called for by Eq. (51), the result is a double-humped resonance curve as shown in Fig. 38 for large values of k. This causes frequencies near resonance to be discriminated against in favor of frequencies at the peaks on either side of resonance. It will be observed that this is the opposite from the type of response that a simple series circuit produces. Hence by combining a double-humped curve such as shown in Fig. 38 with a series-resonant circuit having a suitable Q, a much more uniform over-all response is

obtained. It has been discovered[1] that, when the Q's of the primary and secondary circuits are the same and are high enough to give a pronounced double-humped action, *combining the resulting response with the response of a simple resonant circuit having one-half the Q of the individual resonant circuits in the coupled combination gives a substantially flat-topped band-pass characteristic over the entire pass band irrespective of the coefficient of coupling.* This action is clearly brought out by the curves of Fig. 43. This fact makes it possible to devise a system of circuits in which the band

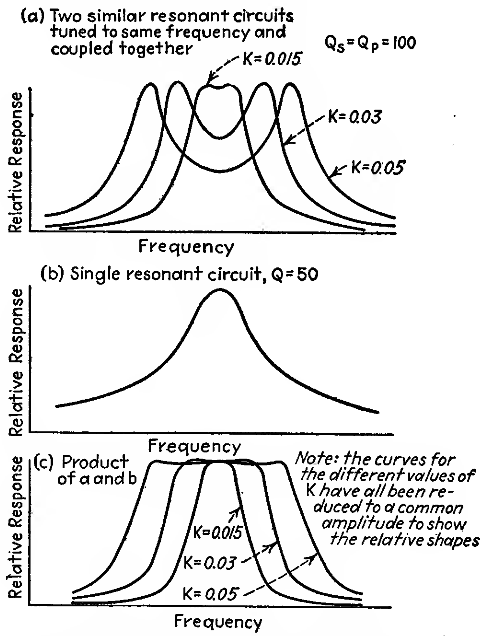


FIG. 43.—Diagram illustrating how a band-pass action can be realized by combining the double-humped curve of a pair of coupled circuits with a simple resonance curve. By using proper proportions the band-pass effect is realized irrespective of the coefficient of coupling between the pair of coupled circuits.

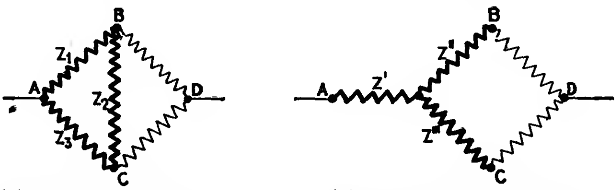
width can be varied by changing the coupling without the necessity of making a simultaneous adjustment of the circuit losses.

20. The Analysis of Complex Circuits.—Complex networks made up of combinations of inductances, resistances, and condensers are most satisfactorily handled by being reduced to simpler equivalent networks. This simplification is carried out by adding impedances that are in series to obtain an equivalent impedance of the series combination and by adding the admittances of parallel impedances in order to obtain an equivalent admittance of the parallel combination. The procedure is exactly the

[1] See Harold A. Wheeler and J. Kelly Johnson, High-fidelity Receivers with Expanding Selectors, *Proc. I.R.E.*, vol. 23, p. 594, June, 1935.

same as used in the analysis of low-frequency alternating-current circuits and need not be considered in detail at this time.

The Δ-Y Transformation.—Bridge circuits of the type shown in Fig. 44*a* cannot be simplified by the usual method of combining impedances and admittances; they are most easily analyzed by replacing one of the Δ's of impedances in the circuit by an equivalent Y of impedances. Thus the impedance triangle $Z_1Z_2Z_3$ of Fig. 44*a*, which cannot be calculated by combining networks in series and parallel, can be replaced by the Y of impedances shown at Fig. 44*b*, which can be made exactly equivalent to the original network as far as the terminals *A*, *B*, and *C* are concerned; that is, the currents and voltages at *A*, *B*, and *C* are the same with the equivalent Y of impedances as in the original network.[1]



(a) Circuit with an Impedance Triangle ABC

(b) Electrically Equivalent Circuit in Which the Delta of Impedances Has Been Replaced by an Equivalent Y

FIG. 44.—Circuits showing how an impedance triangle *ABC* can be replaced by an equivalent Y of impedances.

The Y that is equivalent to a given Δ can be obtained by giving the impedances composing the Y the following values:

$$\left.\begin{aligned} Z' &= \frac{Z_1Z_3}{Z_1 + Z_2 + Z_3} \\ Z'' &= \frac{Z_1Z_2}{Z_1 + Z_2 + Z_3} \\ Z''' &= \frac{Z_2Z_3}{Z_1 + Z_2 + Z_3} \end{aligned}\right\} \tag{52}$$

The notation is given in Fig. 44.

It is also possible to analyze complicated networks, including bridge circuits such as shown in Fig. 44*a*, by means of Kirchhoff's laws, which when applied to alternating-current circuits state: (1) the vector sum of all currents flowing into a junction is zero, and (2) the vector sum of the voltages acting around any closed loop is equal to the vector sum of the voltage drops around the loop. The number of independent equations

[1] This transformation was originated by A. E. Kennelly and published in his paper, The Equivalence of Triangles and Three-pointed Stars in Conducting Networks, *Elec. World and Eng.*, vol. 34, p. 413, 1899. It will be noted that a Δ of impedances is the same thing as a π section and that a Y of impedances can be drawn as a T-section. Hence the Δ-Y transformation is equivalent to a transformation from a π to a T-section.

that can be written according to law 1 is one less than the number of junctions because the equation for the last junction can be obtained by combining the other current equations. The number of these current equations plus the number of voltage equations for the independent loops always equals the number of unknown currents, thus enabling these currents to be determined by a simultaneous solution. The use of Kirchhoff's laws has the disadvantage of requiring the simultaneous solution of a number of vector equations, each of which may be rather complicated; this makes the computations long and increases the likelihood of errors. It is ordinarily preferable to solve a network by combining impedances and admittances rather than by Kirchhoff's laws, and this can always be done except in some circuits involving mutual inductance.

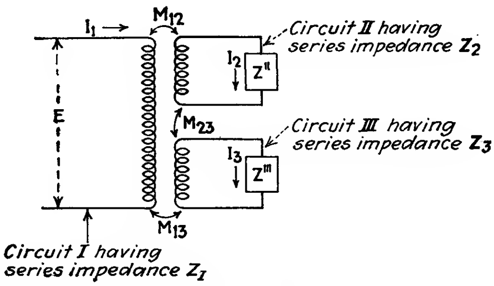


FIG. 45.—A complex coupled circuit involving three mutual inductances.

Circuits Containing Two or More Mutual Inductances.—Circuits involving complex arrangements of mutual inductance must ordinarily be analyzed by means of Kirchhoff's laws. When these laws are applied to coupled circuits, it must be remembered that every current flowing through an inductance induces a voltage in every other coupled inductance. Thus, in Fig. 45, I_1 induces voltages in circuits 2 and 3, I_2 induces voltages in circuits 1 and 3, and I_3 induces voltages in circuits 1 and 2. These induced voltages have a magnitude $-j\omega MI$ where M is the mutual inductance and I the current involved. The way in which the analysis of complex circuits with mutual inductance is carried out can be seen from the following three equations which determine the behavior of the circuit of Fig. 45:

$$\begin{aligned} E - (j\omega M_{12}I_2 + j\omega M_{13}I_3) &= Z_1I_1 \\ (-j\omega M_{12}I_1 - j\omega M_{23}I_3) &= Z_2I_2 \\ (-j\omega M_{13}I_1 - j\omega M_{23}I_2) &= Z_3I_3 \end{aligned}$$

In each of these equations the term in parenthesis represents the induced voltages, and it will be noted that each equation states that the sum total of voltages acting in a circuit is equal to the voltage drop resulting from the circuit current flowing through the circuit impedance Z_1, Z_2,

and Z_3, as the case may be. Simultaneous solution of these equations will give expressions for each of the three currents. The various mutual inductances M_{12}, M_{23}, and M_{13} involved in Fig. 45 may be either plus or minus depending upon the relative direction in which the turns are wound.

Complex circuits may frequently be analyzed qualitatively by inspection. Thus with the circuit of Fig. 46 there will be two frequencies at which the line current will be very great, namely, the frequencies at which one or the other branch is series resonant. In between these two frequencies the reactance of one branch is inductive while that of the other is capacitive, so that there is some frequency in this intermediate range where parallel resonance exists and the line current is extremely small.

When several voltages are simultaneously applied to a circuit, the simplest procedure is to determine the currents and voltages resulting from each component of the applied potential just as though this component existed alone, and then to add the separate effects of the different parts of the applied voltage. This procedure can be followed both when the applied force is composed of components having different frequencies and when there are several voltages of either the same or different frequencies applied at different points in the network.

21. Thévenin's Theorem.—*According to Thévenin's theorem, any linear network containing one or more sources of voltage and having two terminals*

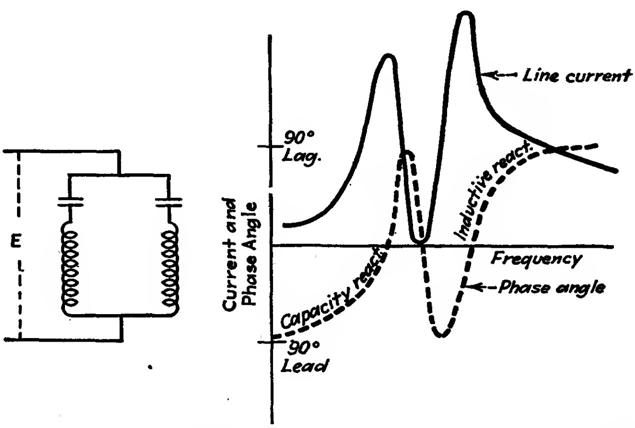


FIG. 46.—Complicated parallel circuit together with sketch showing the current entering the circuit when a constant voltage of varying frequency is applied.

behaves, insofar as a load impedance connected across these terminals is concerned, as though the network and its generators were equivalent to a simple generator having an internal impedance Z and a generated voltage E, where E is the voltage that appears across the terminals when no load impedance is connected and Z is the impedance that is measured between the ter-

minals when all sources of voltage in the network are short-circuited.[1] This theorem means that any network and its generators, represented schematically by the block in Fig. 47*a*, can be replaced by the equivalent circuit shown in Fig. 47*b*. The only limitation to the validity of Thévenin's theorem encountered in ordinary practice is that the circuit elements of the network must be linear, *i.e.*, the voltage developed must always be proportional to current.

Thévenin's theorem offers a very powerful means of simplifying networks, particularly when a load impedance is connected across the output terminals of a complicated network. This theorem will be used repeatedly to assist in the analysis of vacuum-tube amplifier circuits.

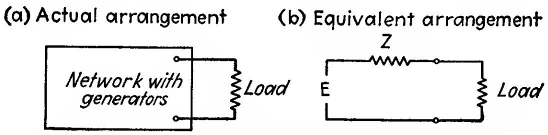


FIG. 47.—Schematic diagram illustrating how Thévenin's theorem can be used to simplify a complicated network containing generators.

Problems

1. Check one or two of the current and phase-angle curves of Fig. 28 by calculating and replotting, using Eq. (27) and the universal resonance curve.

2. A variable condenser having a maximum capacity of 350 $\mu\mu$f, and a minimum capacity of 20 $\mu\mu$f is used for tuning in a broadcast receiver:

a. What size inductance coil is required to make the lowest frequency 530 kc, assuming the coil and associated wiring have a capacity of 20 $\mu\mu$f? Also calculate the exact tuning range with the coil selected.

b. It is also desired to cover the short-wave bands by the use of additional coils to which the condenser can be switched. Assuming these coils and wiring also have 20 $\mu\mu$f of capacity, (1) determine the number of coils required to cover frequencies up to 30,000 kc, (2) specify a suitable inductance value for each coil, and (3) calculate the exact tuning range for each coil chosen.

3. Assume that a series-resonant circuit employs the coil of No. 24 wire shown in Fig. 18 and that the tuning condenser has a power factor of 0.0001. Calculate and plot the current that flows for 1 volt applied when the resonant frequency is 1000 kc, carrying the curve to 50 kc on each side of resonance. In making out these calculations, use the universal resonance curve and Eq. (27).

4. *a.* Using a series circuit employing the same coil as in Prob. 3, and assuming that the losses in the tuning condenser are negligible, calculate and plot upon a common coordinate system the voltage across the condenser as a function of frequency when the applied voltage is 1 mv and when the resonance frequency of the circuit is 600, 1000, and 1400 kc. Carry the curves to 50 kc on each side of resonance, and make use of the universal resonance curve and Eq. (28).

[1] When the sources of energy in the network are constant-current generators instead of constant-voltage generators, the internal impedance Z is the impedance observed between the terminals when all constant-current generators are *open-circuited.* This is due to the fact that a constant-current generator is equivalent to an infinite voltage source having an infinite internal impedance, so that short-circuiting the ultimate source of voltage of the constant-current generator still leaves an infinite impedance in the circuit.

b. From the curves obtained in (*a*), discuss: (1) the effect of the resonant frequency upon the ability of the circuit to discriminate against frequencies that are 50 kc off resonance; (2) the tendency to distort a modulated wave having side-bands extending 5000 cycles on either side of the carrier (corresponding to ordinary broadcast signals), when the circuit is tuned to different carrier frequencies in the broadcast range of frequencies (550 to 1500 kc); (3) the relative response when tuned to different frequencies in the broadcast range.

5. *a*. What is the highest effective Q which a tuned circuit may have when it must respond to a band of frequencies 10,000 cycles wide with a response always at least 70 per cent of the response at resonance, assuming carrier frequencies of 50 kc, 500 kc, 5000 kc, and 50,000 kc.

b. Discuss the significance of the results in Prob. 5*a* with regard to the reception of radio-telephone signals.

6. Check one or two of the impedance and phase-angle curves of Fig. 30 by calculating and replotting, using Eq. (32) and the universal resonance curve.

7. The coil of No. 24 wire in Fig. 18 is tuned to resonance at 800 kc with a condenser of negligible loss. Calculate and plot as a function of frequency: (*a*) magnitude and phase angle of parallel impedance; (*b*) line current, and current in each branch, when the applied potential is 10 volts; (*c*) reactance and resistance components of the impedance of (*a*). Carry these calculations up to 40 kc on either side of resonance, and make use of Eqs. (31), (32), and (33) and the universal resonance curve.

8. A particular coil having an inductance of 0.18 mh has a distributed capacity of 10 $\mu\mu$f. Calculate and plot the factors by which the actual inductance, actual resistance, and actual Q must be multiplied to give the apparent or observed values, carrying out these calculations to a frequency closely approaching the point where the coil is in resonance with the distributed capacity. (Note that the results at frequencies in the vicinity of resonance are only approximate because the distributed capacity is assumed to be lumped.)

9. Sketch the types of current and voltage distribution that would be expected on a line 1⅞ wave lengths long when the receiver is: (*a*) an open circuit, (*b*) a short circuit, (*c*) a resistance equal to the characteristic impedance.

10. Two identical coils each having $Q = 100$ and an inductance of 200 μh are coupled together with a mutual inductance of 50 μh. If the secondary coil is short-circuited, calculate: (*a*) the coupled resistance and reactance at a frequency of 600 kc, (*b*) the total resistance and reactance of the primary circuit, (*c*) the effective Q of the primary circuit including the effect of the coupled impedance.

11. Describe a procedure for experimentally determining the coefficient of coupling between a coil and its shield can, assuming that the shield has negligible resistance.

12. Discuss the effect that a copper shield has on the inductance and resistance of a coil as compared with the effect of a shield of identical dimensions but having a higher resistivity, such as a brass shield.

13. In the circuit shown in Fig. 37, assuming $\omega L_p << R_p$, calculate and plot curves up to 40 kc on either side of resonance showing: (*a*) magnitude, phase angle, resistance component and reactance component of impedance coupled into the primary; (*b*) magnitude of total primary impedance; (*c*) magnitude of primary current; (*d*) magnitude of voltage induced in the secondary; (*e*) voltage across the secondary condenser. In drawing these curves use one set of axes for (*a*) and another for the curves in (*b*) to (*e*).

14. Recalculate the curves for $k = 0.03$ of Fig. 38 by going through the following steps:

a. Calculate and plot the resistance and reactance components of the coupled impedance.

b. Calculate and plot the resistance and reactance components of the primary circuit when the secondary is removed.

c. Add (*a*) and (*b*) to obtain the curve of total primary-circuit resistance and reactance, and convert the results into curves giving the magnitude and phase of the primary impedance.

d. Calculate and plot magnitude of primary current and of secondary induced voltage.

e. Calculate and plot the secondary-current curve.

Carry these calculations to 40 kc off resonance, and plot the various sets of curves above one another.

15. Two identical tuned circuits with $L = 180$ μh and $Q = 100$ are coupled together and both tuned to a resonant frequency of 800 kc.

a. Calculate and plot the secondary current at resonance as a function of mutual inductance when 1 volt is applied in series with the primary.

b. For couplings greater than the critical values calculate and plot width between humps of secondary current as a function of mutual inductance.

16. Calculate and plot the coefficient of coupling as a function of frequency in the circuit of Fig. 40*e* when $L_1 = L_2 = 180$ μh, $C_1 = C_2 = 169$ $\mu\mu$f, $C_m = 3380$ $\mu\mu$f, and $L_m = 9$ μh. Carry the curves over the frequency range of 700 to 1300 kc.

17. Derive formulas for the equivalent mutual inductance of the circuits of Figs. 40*a* and 40*f*.

18. A particular band-pass filter for use as an intermediate-frequency transformer in a radio receiver must have a band width of 7000 cycles centering about a frequency of 456 kc. If the primary and secondary inductances are both 2 mh:

a. Specify the tuning capacities, the proper coefficient of coupling, and the proper circuit Q.

b. Calculate the shape of the response curve obtained when 1 mv is considered as applied in series with the primary circuit.

19. Signals in the frequency range of 550 to 1500 kc are to be tuned by means of a band-pass filter for which a constant band width of 8000 cycles would be ideal. The circuits are assumed to have $Q = 100$ over this frequency range. If the adjustment is such that $k = 0.013$ at 550 kc, discuss how the width and shape of the pass band will vary with resonant frequency when the tuning is obtained by varying the primary and secondary condensers simultaneously and when the coupling is (*a*) inductive as shown at Fig. 42*a* and *b*, (*b*) capacitive as shown at Fig. 42*c*. Illustrate the discussion with the aid of sketches showing types of response curves to be expected under various conditions.

20. Assume that the circuit of Fig. 44*a* represents a bridge not perfectly balanced and that arm BC represents the galvanometer, while a direct-current voltage of 10 volts is applied to AD. If $Z_1 = 1000$ ohms, $Z_3 = 10{,}000$ ohms, $Z_2 = 5000$ ohms, $BD = 1000$ ohms, and $DC = 9000$ ohms, calculate the current through Z_2, using both the Δ-Y transformation and Thévenin's theorem.

21. The two inductances in the circuit of Fig. 46 are both 10 mh, while the capacities are 0.005 and 0.003 mf, respectively. Neglecting the circuit resistance, calculate and plot the following as a function of frequency:

a. Reactance of the individual branches.

b. Current in the individual branches.

c. Total reactance of the parallel combination.

d. Line current for the parallel combination.

22. By means of Thévenin's theorem, derive a formula for the secondary current in a circuit of the type shown in Fig. 37.

CHAPTER IV

FUNDAMENTAL PROPERTIES OF VACUUM TUBES

22. Vacuum Tubes.—Vacuum tubes are used to generate the radio-frequency power required by a radio transmitter, to control the energy radiated, to amplify the weak radio-frequency signals obtained at the receiver, to rectify the signal, to amplify this rectified signal, and so on. The amplifying properties of vacuum tubes have also made possible the long-distance telephone, the talking picture, the modern phonograph, public-address systems, television, and so on. Altogether it may be said that the vacuum tube is probably the most important single piece of equipment that has come into electrical engineering since the beginning of the century.

In its usual form the vacuum tube includes a cathode capable of emitting electrons when heated, an anode (often called the plate) that attracts the electrons emitted from the cathode, and some means of controlling the flow of electrons from the cathode to the anode. These electrodes are enclosed in a gas-tight space, evacuated to a degree which usually represents the highest vacuum that it is practicable to obtain, in order to permit the electrons to flow unimpeded from cathode to anode. There are, in addition, modified forms of vacuum tubes, such as those having only a cathode and anode, tubes with more than three electrodes, and tubes into which small quantities of gas have been intentionally introduced to produce certain operating characteristics.

Tubes are classified as diodes, triodes, tetrodes, pentodes, etc., according to whether there are two, three, four, five, etc., electrodes present. Thus a tube with only cathode and anode is a diode, while the addition of a control electrode (a grid) converts it into a triode.

23. Electrons and Ions.—Electrons can be considered as minute negatively charged particles which are constituents of all matter. They have a mass of 9×10^{-28} gram, which is $1/1840$ that of a hydrogen atom, and a charge of 1.59×10^{-19} coulomb; they are also always identical irrespective of the source from which derived.[1] Atoms are composed of one or more such electrons associated with a much heavier nucleus which has a positive charge equal to the number of the negatively charged

[1] Recent studies have shown that, in addition to the usual properties of a moving charged body, electrons in motion possess wave characteristics, which, however, are not of practical importance as far as vacuum-tube technique is concerned.

electrons contained in the atom, so that an atom with its full quota of electrons is electrically neutral. The differences between chemical elements arise from differences in the nucleus and in the number of associated electrons but not from variations in the character of electrons, which are always the same.

Positive ions represent atoms or molecules that have lost one or more electrons and so have become charged bodies having the weight of the atom or molecule concerned and a charge equal to that of the lost electrons. Unlike electrons, positive ions are not all alike and may differ in charge or weight, or both. They are much heavier than electrons and resemble the molecule or atom from which derived. Ions are designated according to their origin, such as mercury ions, hydrogen ions, etc.

Electrons and ions are produced by separating the constituent parts of the atom or molecule in such a way as to produce molecules that are deficient in electrons, and free electrons. There are a number of ways in which this separation may be accomplished. Thus in a gas, when a swiftly moving ion or electron collides with a molecule, the impact may be sufficiently intense to knock one or more electrons out of the molecule, producing one or more free electrons and leaving a positive ion. This method of producing ions and electrons is known as *ionization by collision* and occurs in all vacuum tubes in which gas is present. Again, if a solid body is sufficiently hot, some of the electrons that make up the solid material will escape from its surface into the surrounding space, thus giving free electrons which are said to be obtained by *thermionic emission.* When ultra-violet light or x-rays strike a solid body or a gas, electrons will be emitted even at normal temperatures; with certain substances, notably potassium, caesium, and other alkaline earths, visible light will cause electrons to be emitted into the space surrounding the material. Electrons obtained in this way by the use of light radiation are said to result from the *photoelectric effect.* Electrons can also be obtained from solid materials as a result of impact of rapidly moving electrons or ions, which can knock electrons out of a solid body when striking with sufficient velocity. Electrons obtained in this way are said to result from *secondary electron emission* because it is necessary to have some primary source of electrons (or ions) before the secondary electron emission can be obtained. Finally, it is possible to pull electrons directly out of solid substances by an intense electrostatic field at the surface of the material.

24. Motions of Electrons and Ions.—Electrons and ions are charged particles and so have forces exerted upon them by an electrostatic field in the same way that other charged bodies do. The electrons, being negatively charged, tend to travel toward the positive or anode electrode while the positively charged ions travel in the other direction toward the negative or cathode electrode. The force exerted upon a charged

particle by an electrostatic field is proportional to the product of the charge e of the particle and the voltage gradient G of the electrostatic field. Expressed in the form of an equation this relation is

$$\begin{aligned}\text{Force in dynes} &= \left(\begin{matrix}\text{gradient } G \text{ in}\\ \text{volts per centimeter}\end{matrix}\right)(\text{charge } e \text{ in coulombs})10^7 \qquad (53a)\\ &= Ge \times 10^7\end{aligned}$$

This force upon the ion or electron is in the direction of the electrostatic-flux lines at the point where the charge is located and acts toward or away from the positive terminal, depending on whether a negative or positive charge, respectively, is involved. The force that the field exerts on the charged particle causes an acceleration in the direction of the field at a rate that can be calculated by the ordinary laws of mechanics where the velocity does not approach that of light. That is

$$\text{Acceleration in centimeters per second per second} = \frac{\text{force in dynes}}{\text{weight in grams}} \qquad (53b)$$

It is to be noted that, as long as there is an electrostatic field acting in the direction in which the electron or ion is moving, the velocity will increase because an acceleration will be produced even when the field is weak.

The velocity that an electron or ion acquires in being acted upon by an electrostatic field can be calculated in the usual way by the laws of mechanics provided the velocities involved are well below the velocity of light. The amount of energy that a body with a charge of e coulombs gains in traveling between two points between which a difference of potential of V volts exists is equal to Ve joules, as can readily be proved by integrating Eq. (53*a*). This energy is all converted into kinetic energy of motion, so that the velocity can be obtained from the formula

$$\left.\begin{aligned}&\text{Kinetic energy in ergs} = Ve10^7 = \tfrac{1}{2}mv^2\\ \text{or}\quad &\\ &v = \text{velocity in centimeters per second corre-}\\ &\quad\text{sponding to voltage } V = \sqrt{\frac{2Ve10^7}{m}}\end{aligned}\right\} \qquad (53c)$$

The velocity with which electrons and ions move in even moderate-strength fields is very great. Thus an electron in falling through a potential difference of 10 volts will gain a velocity of 1160 miles per second. These velocities are so great that from a practical point of view it is much simpler to express the velocity in terms of the difference of potential through which the electron (or ion) has fallen in gaining its speed. Thus, when it is stated that an electron has a velocity of 10 volts,

one means the velocity that an electron would obtain in falling through a voltage drop of 10 volts. Since the velocity that a charged particle gains in falling through a difference of voltage is inversely proportional to the square root of the weight of the particle, the velocity represented by a given voltage will depend upon the weight of the charged body and will be much greater with electrons than ions, particularly the heavy ions. Thus a mercury ion having a charge equal to that of an electron will move less than one six-hundredth as fast as an electron in the same electrostatic field.

Effect of a Magnetic Field on a Moving Electron.—Since a moving charge represents an electrical current, an ion or electron in motion has all the properties of an electrical current. Most important of these, so far as vacuum tubes are concerned, is the force which a magnetic field at right angles to the direction of current flow (*i.e.*, line of travel of electron or ion) exerts on a moving charge and which is exerted at right angles to both the magnetic field and the line of flow of current (*i.e.*, charge). Since the current represented by a moving charge is equal to the product of the charge and its velocity, a charge of Q coulombs moving with a velocity v represents a current Qv, and the force that is exerted upon this moving charge by a magnetic field having an intensity of H gausses is $HQv/10$ dynes in a direction at right angles to the motion of the charge. In the case of an electron which has a charge of e, the force is accordingly

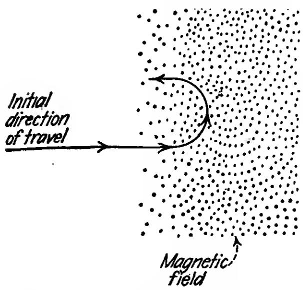


FIG. 48.—Path of electron in magnetic field.

$$\text{Force in dynes} = \frac{Hev}{10} \tag{54}$$

An electron shot with high velocity into a magnetic field will follow a path similar to that shown in Fig. 48. The acceleration that the forces of the magnetic field produce on such an electron are always at right angles to the direction in which the electron is traveling and will result in a circular path when the field is uniform. The radius of this circle is smaller the greater the strength of the magnetic field and the more slowly the electron is moving through the field. When electrons or ions are under the simultaneous influence of both electrostatic and electromagnetic fields, the resultant force that acts upon the moving charge at any instant is the vector sum of the two component forces.

It can be shown that an electron or ion that is being accelerated will radiate a certain amount of energy in the form of electromagnetic waves. Thus, when an electron moving at high velocity is suddenly stopped by impact against a metallic surface, x-rays of a wave length depending

upon the velocity of the electron will be produced. An alternating acceleration, such as can be produced by an alternating voltage, will cause an electron to radiate electromagnetic waves of the type employed in radio communication. It is this property of electrons to radiate energy when accelerated that accounts for many of the effects produced on radio waves by ionization in the upper atmosphere.

25. Thermionic Emission of Electrons.—The electrical conductivity of metals is the result of electrons within the material which for the moment are not definitely attached to any particular molecule. These free electrons move about inside the conductor with a velocity that increases with temperature and they exert a pressure just as does an ordinary gas. This pressure does not cause electrons to escape from the metal into the neighboring space, however, because there are attractive forces at the surface that tend to keep the electrons in the substance and these forces are much greater in magnitude than the pressure of the "electron gas."

In order to escape from the surface of a conductor, an electron must therefore do a certain amount of work to overcome the surface forces, and this energy can be obtained only from the kinetic energy possessed by the electron as a result of its motion. *Unless the kinetic energy of an electron exceeds the work that the electron must perform to overcome the surface forces of the conductor, the electron cannot escape.* For all known substances this energy which a free electron must have in order to be emitted from the material is so related to the kinetic energy possessed by the electrons that practically none escape at ordinary room temperatures. It is only at high temperatures, where the average kinetic energy possessed by the free electrons is large, that an appreciable number will have sufficient kinetic energy to escape through the surface of the material.

The process of electron emission from a solid substance is very similar to the evaporation of vapor from the surface of a liquid. In the case of the vapor the evaporated molecules represent molecules that obtained sufficient kinetic energy to overcome the restraining forces at the surface of the liquid, and the number of such molecules increases rapidly as the temperature is raised. The thermionic emission of electrons from hot bodies is seen to represent the same process and can be considered as an evaporation of electrons in which the energy the electron must give up in escaping corresponds to the latent heat of vaporization of a liquid.

The number of electrons evaporated per unit area of emitting surface is related to the absolute temperature T of the emitting material and a quantity b that is a measure of the work an electron must perform in escaping through the surface, according to the equation

$$I = AT^2\epsilon^{-\frac{b}{T}} \tag{55}$$

where I is the electron current in amperes per square centimeter and A is a constant, the value of which may vary with the type of emitter. The temperature at which the electron current becomes appreciable is determined almost solely by the quantity b, which is accordingly the most important characteristic of an electron-emitting material. The emission is very sensitive to changes in b and the temperature, for these quantities appear in the exponent of Eq. (55); slight variations in the value of either change the magnitude of the exponential term enormously. The value of A is therefore of secondary importance, for the effects of wide variations in A can be compensated for by small temperature

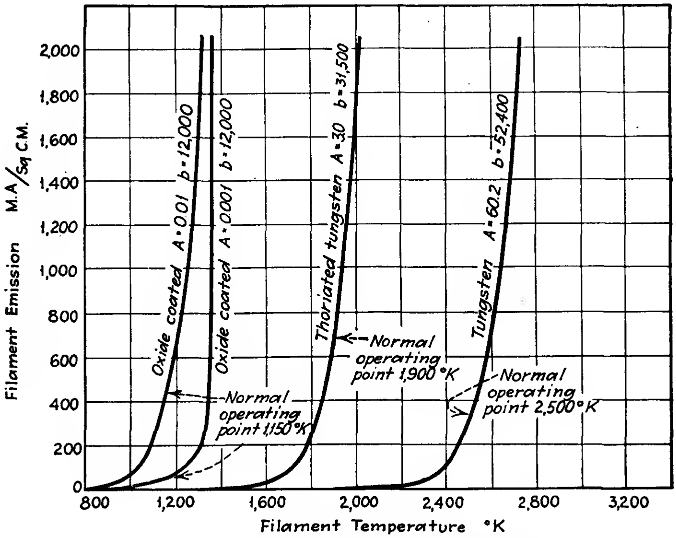


FIG. 49.—Variation of electron emission with absolute temperature for the three principal types of electron emitters, showing the extreme sensitiveness of the emission to temperature and to the value of the constant b.

changes. The relationship between electron emission per unit area of emitter and the emitter temperature is given in Fig. 49 for several typical electron-emitting materials commonly used in vacuum tubes and shows how extremely sensitive electron emission is to temperature and to the value of the work function b, while being much less dependent on A.

26. Electron-emitting Materials.[1]—The properties of matter are such that thermionic emission of electrons does not become appreciable until temperatures of the order of 700 to 2500°K. are reached, the exact value depending upon the material involved. The high temperatures

[1] This section is a very brief summary of the principal characteristics of the types of electron emitters employed in commercial vacuum tubes. For further information

required limit the number of substances suitable for use as thermionic electron emitters to a very few, of which tungsten, thoriated-tungsten, and oxide-coated emitters are the only ones commonly used in vacuum tubes.

Tungsten.—While a relatively poor emitter, tungsten can be operated at temperatures so high as to give good emission in spite of the large value of work function b which it possesses. Tungsten is extensively used as the electron-emitting cathode of high-power vacuum tubes because of its durability under the exacting conditions encountered in such tubes. The essential characteristics of tungsten emitters are shown in Table IV.

Thoriated Tungsten.—Thoriated-tungsten emitters have an electron emission that is many thousand times that of pure tungsten operated at the same temperature, and begin to emit an appreciable amount at temperatures 500° to 600°K. lower than for pure tungsten. This increased electron emission is the result of a layer of thorium one molecule deep that forms on the surface of the thoriated tungsten and reduces the work that an electron must do to escape.[1] Thoriated-tungsten emitters consist of tungsten containing a reducing agent (ordinarily carbon) and a small quantity (1 to 2 per cent) of thorium oxide. Such cathodes must be activated by heating the emitter to approximately 2600° to 2800°K. for one or two minutes, which is called flashing, and then glowing for some minutes at an activating temperature of 2100° to 2200°K. The flashing raises the temperature to the point where the impregnated carbon reduces some of the thorium oxide to metallic thorium, and the subsequent treatment at the activating temperature allows this thorium to diffuse to the surface, where it forms a layer one molecule deep which is the seat of the high electron emission. During operation of a thoriated-tungsten emitter, thorium is being continuously evaporated from the surface layer but is replenished by diffusion from the interior of the tungsten. The important characteristics of thoriated-tungsten emitters are given in Table IV.

The performance of thoriated-tungsten emitters can be improved by carbonizing the tungsten. This is done by glowing the filament at a temperature exceeding 1600°K. in a hydrocarbon atmosphere. Hydrocarbon molecules striking the hot filament are then dissociated and the free carbon thus produced combines with the tungsten to form tungsten

---

consult Saul Dushman, Thermionic Emission, *Rev. Modern Phys.*, vol. 2, p. 381, October, 1930; Saul Dushman, Electron Emission, *Elec. Eng.*, vol. 53, p. 1054, July, 1934; E. L. Chaffee, "Theory of Thermionic Vacuum Tubes," Chap. IV, McGraw-Hill Book Company, Inc.

[1] For further information relative to thoriated-tungsten emitters, the reader is referred to Irving Langmuir, The Electron Emission from Thoriated-tungsten Filaments, *Phys. Rev.*, vol. 22, p. 357, 1923.

carbide. The mono-molecular thorium layer clings much more firmly to this carbide surface than to pure tungsten, permitting operation at higher temperatures than is otherwise possible without evaporating off the thorium surface. This increases the thermionic emission and also the rate at which thorium diffuses to the surface to become available for maintaining the mono-molecular surface layer.

Oxide-coated Emitter.—The oxide-coated emitter consists of a mixture of barium and strontium oxides coated on the surface of a suitable metal, such as a nickel or a platinum alloy, and when suitably activated will emit large numbers of electrons at temperatures of the order of 1150°K. The electron emission of emitters of the oxide-coated type appears to arise from a layer of alkaline-earth metal, *i.e.*, metallic barium and strontium, which forms on the surface of the oxide coating; emission is at a maximum when this layer covers the entire surface of the oxide to a depth of approximately one molecule. During operation of the emitter at its normal working temperature the surface metal that is evaporated is replenished by diffusion of additional molecules from the interior of the oxide coating.

Oxide-coated emitters when first formed must be activated (or "formed") before the full thermionic activity is realized. One method consists in glowing the emitter for several minutes at a temperature of approximately 1500°K. and then applying a strong positive electrostatic potential gradient for a period of from 2 to 30 min. During this period the electron emission increases, and, when it has reached a rather large value, the electrostatic field and emitter temperature are readjusted to lower values, which are maintained for an additional period. The exact procedure for activating oxide-coated emitters is in the nature of an art which varies greatly under different circumstances but which commonly follows the broad outline that has been given. The object of the activation process is to break down a fraction of the oxide coating into the alkaline-earth metal and to use this metal to build up a layer of metallic molecules on the surface of the oxide. The high temperature that is used at the start of the activation process causes some of the oxide coating to dissociate into ions, and the positive or metal ions so produced migrate inward when electrons are drawn from the emitter and are transformed into metal molecules at the interface between core and oxide. These molecules then diffuse to the surface through the porous oxide coating. The increase in electron emission which takes place during the forming process is caused by the gradual building up of the surface layer of alkaline-earth metal. The high thermionic activity of oxide-coated cathodes appears to be caused by this surface layer becoming positively charged as a result of losing electrons and then acting as a positively charged screen or grid which covers the surface of the oxide coating and

helps pull electrons from its surface.[1] The essential characteristics of properly formed oxide-coated emitters are to be found in Table IV.

TABLE IV.—CHARACTERISTICS OF ELECTRON EMITTERS

Type of emitter	Constants in equation $i = AT^2\epsilon^{-\frac{b}{T}}$		Normal operating temperature, °K.	Efficiency at operating temperature in milliamperes emission per watt of heating power*
	A*	b*		
Tungsten	60.2	52,400	2450 to 2600	3 to 15
Thoriated tungsten	3.0	31,500	1900	62.5
Oxide coated	0.01 to 0.001	12,000	1100 to 1170	50 to 125

* The values of A, b, and efficiency vary greatly with different oxide-coated cathodes and to a lesser extent with thoriated tungsten. The values given in this table are typical for good emitters but do not necessarily apply to every oxide-coated and thoriated-tungsten cathode.

Power Required to Heat Emitter.—The electron-emitting cathodes of vacuum tubes are heated electrically, either by forming the emitter into a filament that is raised to the necessary temperature by the passage through it of a suitable current, or by using a cylindrical cathode that is heated either by conduction or radiation from an internal heater consisting of an incandescent tungsten filament. Filament cathodes may be of the tungsten, thoriated-tungsten, or oxide-coated type, while heater-type cathodes always employ an oxide-coated emitter because of the impossibility of obtaining by indirect heating the high temperatures required by other emitters.

Most of the power required to maintain the cathode at a given temperature represents energy sent out from the cathode in the form of radiant heat. A small amount of power is lost by heat energy conducted away from the cathode along the support and lead wires, but other than this the loss of heat by conduction is negligible because the cathode is in a very high vacuum. When an object is at a temperature considerably higher than surrounding objects, the heat energy radiated is proportional

[1] The exact mechanism of electron emission from an oxide-coated electrode is not yet completely understood, although considerable progress in this direction has been made in recent years. An easily readable qualitative discussion is given by J. A. Becker, The Role of Barium in Vacuum Tubes, *Bell Lab. Record*, vol. 9, p. 54, October, 1930. For more comprehensive and scientific discussions, see Saul Dushman, Thermionic Emission, *Rev. Modern Phys.*, vol. 2, p. 381, October, 1930; J. A. Becker, Phenomena in Oxide-coated Filaments, *Phys. Rev.*, vol. 34, p. 1323, November, 1929; J. A. Becker and R. W. Sears, Phenomena in Oxide-coated Filaments II. Origin of Enhanced Emission, *Phys. Rev.*, vol. 38, p. 2193, December, 1931; J. A. Becker, Thermionic Electron Emission, *Rev. Modern Phys.*, vol. 7, p. 95, April, 1935 (also reprinted in *Bell System Tech. Jour.*, vol. 14, p. 413, July, 1935).

to the fourth power of the absolute temperature, and so is given by the relation

$$\text{Radiated energy per square centimeter of surface} = KT^4 \qquad (56)$$

where T is the temperature in degrees Kelvin and K is a constant, the exact value of which depends on the surface conditions. As a consequence of this relation the electrical power required to heat the cathode

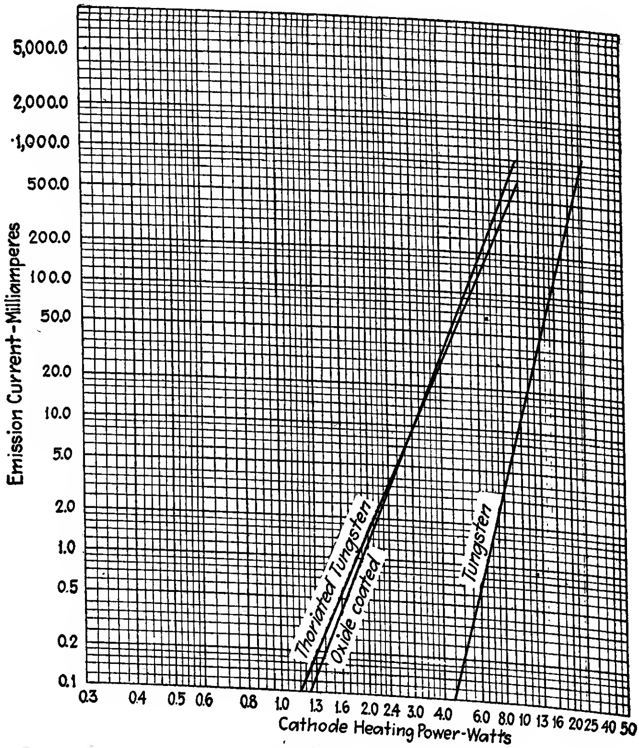


FIG. 50*a*.—Relation between cathode heating power and electron emission for three representative cathodes, plotted on power-emission paper. These curves are straight lines and so can be drawn when only three points are known, and can be extrapolated as far as desired.

increases very rapidly with temperature, and emitters such as those of tungsten, which must operate at high temperature, require considerably more heating power in proportion to electron emission than do the low-temperature emitters of the oxide-coated type. These differences are demonstrated by the column of milliamperes electron emission per watt of cathode heating power given in Table IV for different emitters.

Equation (56) can be substituted in Eq. (55) to give an equation relating electron emission to the cathode heating power. From this

equation the curvilinear system of coordinates shown in Fig. 50*a* can be derived on which the electron emission plots as a straight-line function of the heating power. Coordinate paper of this type,[1] which is known as power-emission paper, is very useful in the comparison of different cathodes because the cathode heating power required to give three values of electron emission, such as 0.1, 1.0, and 10 ma, suffices to determine the

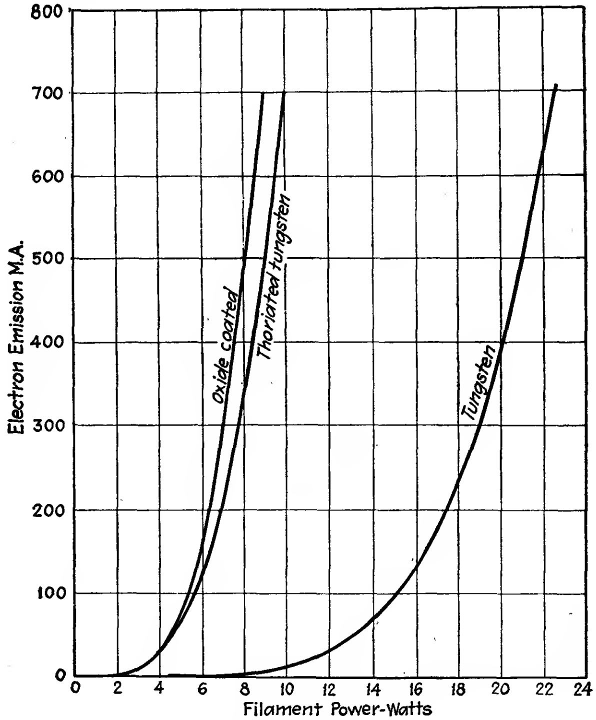


FIG. 50*b*.—Relation between cathode heating power and electron emission for three representative cathodes, plotted in rectangular coordinates. The very sudden rise in these curves as contrasted with the same data plotted on Fig. 50*a* emphasizes the usefulness of the power-emission paper.

straight line on the power-emission paper and to permit accurate extrapolation to emissions that could not be maintained for even a few seconds without damage to the tube. Power-emission curves for the cathodes of Fig. 49 are shown on power-emission paper in Fig. 50*a*; when contrasted with similar curves on rectangular coordinates, as shown in Fig. 50*b*, they illustrate the usefulness of the curvilinear coordinate system in straightening out the plotted emission characteristic.

Operating Temperature and Life.—The temperature at which tungsten emitters are operated is a compromise between high electron emission

[1] This paper is obtainable from Keuffel and Esser, under the name "Power Emission Chart."

per watt of heating power, which means a high operating temperature and a short life, and a low emission per watt of heating power, which goes with a low operating temperature and a long life. The actual temperature selected is the highest value that will give a life expectation of 1000 to 2000 hr. The life is determined by the rate at which tungsten evaporates from the surface of the filament, and, since a given depth of evaporation is less important when the filament is thick, large-diameter filaments can be operated at slightly higher temperatures than thin filaments and so give greater electron emission per watt of heating power.

The life of thoriated-tungsten filaments is determined by the supply of thorium in the tungsten and by the rate at which the thorium is evaporated from the surface layer. Filaments of this type never burn out, as do those of tungsten, except as a result of an accident. The proper operating temperature for thoriated tungsten is very close to 1900°K., at which the total life expectancy is at least several thousand hours of service. If the operation is at a temperature much above 2000°K., the thorium evaporates from the surface layer faster than this layer is replenished by diffusion of thorium from the interior of the tungsten, with the result that the surface is soon denuded of its thorium layer and the electron emission becomes that of pure tungsten. A thoriated-tungsten filament that has been in service for many hours will begin to show low electron emission as a result of exhaustion of the metallic thorium produced in the tungsten at the time of activation. The original electron emission can be restored by flashing the filament for 20 to 30 sec. at three to four times normal voltage and then burning for 30 to 60 min. at an overvoltage of 25 to 40 per cent. This process, which is often termed rejuvenation, reactivates the filament by reducing additional thorium oxide to metallic thorium and reforming the surface layer of thorium; and it can be repeated until the supply of thorium oxide is exhausted.

The life of oxide-coated cathodes is limited by the supply of active electron-emitting material in the cathode. The life of these cathodes is very great, often in excess of 5000 hr., and many tubes with oxide-coated cathodes have been operated continuously for more than three years before the electron emission became seriously reduced. Oxide-coated filaments do not require occasional rejuvenation because the activation process of electrolysis and diffusion goes on in them continuously during use. The operating temperature is a compromise between life and efficiency of emission (*i.e.*, electron emission per watt of heating power) and is not so critical as in the case of tungsten and thoriated-tungsten cathodes.

The life of thoriated-tungsten and oxide-coated cathodes is very adversely affected by the presence of small traces of gas within the tube,

particularly when there is a high anode voltage. This is due to the fact that emitted electrons in flowing to an anode will ionize residual gas by collision, producing positive ions which are attracted toward the cathode and bombard its surface. In the case of thoriated-tungsten emitters, these positive ions strip off the mono-molecular layer of thorium on the tungsten surface that causes the high emissivity. Carbonizing the thoriated-tungsten filament reduces greatly the detrimental action of positive-ion bombardment by causing the thorium molecules to stick more tenaciously to the carbonized surface than to pure tungsten. Carbonizing also permits operation at a higher temperature than is otherwise possible without evaporating off the thorium layer, and this higher temperature increases the rate at which thorium molecules diffuse to the surface to replace the losses caused by gas-ion bombardment. In the case of oxide-coated emitters, bombardment by positive ions having a velocity greater than about 15 to 25 volts will cause the cathode to disintegrate mechanically.

Uses of Different Emitters.—Tungsten, thoriated-tungsten, and oxide-coated cathodes have special fields of usefulness. Oxide-coated emitters have longer life and greater emission per watt of heating power than other types and so are used wherever possible. All heater-type tubes as well as practically all other tubes used in radio receivers have oxide-coated emitters. Oxide-coated emitters are not satisfactory where the anode potential exceeds 500 to 2000 volts because of the detrimental effects of positive-ion bombardment, together with the fact that tubes with oxide-coated cathodes cannot be degassed so thoroughly as other types of tubes. The high temperatures required to accomplish complete degassing will destroy the cathode activity.

Thoriated-tungsten emitters are used in tubes operated at moderate voltages, particularly in the range of anode potentials between 500 and 5000 volts. Such emitters, while less efficient than those of the oxide-coated type, are still much more efficient than tungsten and, when carbonized, will stand up satisfactorily under these potentials provided the tubes are thoroughly degassed at the time of evacuation.

Tungsten emitters will withstand a positive-ion bombardment that would quickly make other types of emitters inactive, but they are relatively inefficient from the point of view of the milliamperes of emission per watt of heating power. Consequently tungsten filaments are used only in tubes where the anode voltage is so high or the gas conditions are such that thoriated-tungsten and oxide-coated cathodes are out of the question.

Velocity of Emission.—The electrons emitted from a hot cathode come out with a velocity that represents the difference between the kinetic energy possessed by the electron just before emission and the energy

that must be given up to escape. Since the energy of the different electrons within the emitter is not the same, the velocity of emission will be different for different electrons, and will commonly range from zero up to over 1 volt. Experiment shows that the velocities of emission are distributed according to Maxwell's law for the distribution of velocities in a gas composed of electrons and having the temperature of the emitting cathode.[1] The average velocity of emission accordingly increases with the cathode temperature, just as does the average velocity of gas molecules.

27. Current Flow in a Two-electrode Tube; Space-charge Effects.—When an electron-emitting cathode is surrounded by a positive anode (*i.e.*, plate electrode) to form a two-electrode vacuum tube (or diode), the relation between the plate current (*i.e.*, the number of emitted electrons that are attracted to the anode) and the plate potential has the character shown in Fig. 51, which gives the results obtained in a typical tube for several cathode temperatures. It is seen that at high anode voltages the electron current is largely independent of anode voltage, being determined primarily by the cathode temperature, while at low anode voltages the current is controlled by the anode voltage and is independent of cathode temperature. When the plate is negative, it repels electrons and the plate current is then zero.

The behavior observed at high anode voltages is a result of the fact that a high anode potential draws the electrons away from the filament as fast as they are emitted, which makes the anode current equal the total electron emission from the cathode. Under these conditions the anode current is given by Eq. (55) and the tube is said to be operating at voltage saturation.[2]

Space-charge Effects.—At low plate (*i.e.*, anode) voltages the anode current is limited by the repelling effect that the negative electrons already in the space between anode and cathode have on the electrons just being emitted from the cathode. The electrons in the interelectrode space constitute a negative space charge (*i.e.*, a negative charge distributed in space), and at any instant the number of electrons that are in transit between electrodes cannot exceed the number that will produce a negative space charge which completely neutralizes the attraction that

[1] See L. H. Germer, The Distribution of Initial Velocities among Thermionic Electrons, *Phys. Rev.*, vol. 25, p. 795, 1925.

[2] The sharpness with which voltage-saturation effects appear differs greatly with the type of emitter. Thus the anode current with cathodes of tungsten or thoriated tungsten has a characteristic such as shown by the dotted lines in Fig. 51, in which the saturation effect is almost complete, while in emitters of the oxide-coated type the saturation effect takes place more gradually, as shown by the solid lines of Fig. 51. The sharpness with which voltage saturation appears is also reduced by the cooling of the ends of the cathode by the support wires, and by the voltage drop along the cathode of filament-type tubes.

the positive plate exerts upon the electrons just leaving the cathode. All electrons in excess of the number necessary to neutralize the effect of the plate voltage are repelled back into the cathode by the negative space charge of the electrons in transit, so that the anode current will be independent of the electron-emitting power of the cathode, provided the cathode is capable of emitting enough electrons to produce a full space charge. When a full space charge is present, the plate current depends

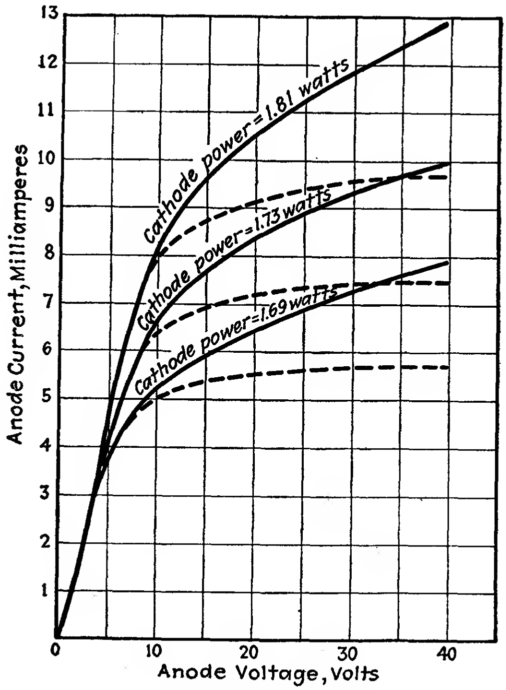


FIG. 51.—Anode current as a function of anode voltage in a two-electrode tube for three cathode temperatures. The solid lines are the characteristics actually obtained using an oxide-coated cathode, while the dotted lines show the type of curve that is given by tungsten and thoriated-tungsten cathodes.

upon the plate voltage, since with higher voltages the electrons travel from cathode to anode more rapidly, making the rate of arrival greater in proportion to the total number in the space between anode and cathode at any instant; furthermore, it takes more space charge to neutralize the effect of the higher voltage. Increasing the plate voltage thus causes the electron flow to increase until a point is reached where the total electron emission of the cathode is being drawn to the plate, after which further increases in anode voltage will produce practically no additional current because of voltage saturation.

The energy that is delivered to the tube by the source of anode voltage is first expended in accelerating the electrons traveling from cathode to anode and so is converted into kinetic energy. When these swiftly moving electrons strike the anode, this kinetic energy is then transformed into heat as a result of the impact and appears at the anode in the form of heat that must be radiated to the walls of the tube.

When the anode current is limited by space charge, the negative charge of the electrons in transit between cathode and plate will be sufficient to give the space in the immediate vicinity of the cathode a slight negative potential with respect to the cathode. The electrons emitted from the cathode are projected out into this negative field with an emission velocity that will vary with different electrons. The negative field next to the cathode causes the emitted electrons to slow down as they move away from the cathode, and those having a low velocity of emission are driven back into the cathode. Only those electrons having the highest velocities of emission will be sent out with sufficient force to penetrate through the negative field near the cathode and reach the region where they are drawn toward the positive plate. The remainder, *i.e.*, those electrons having low emission velocities, will be brought to a stop by the negative field adjacent to the cathode and will fall back into the cathode.

Anode Current When Limited by Space Charge.—When limited by space charge, the anode current received from any portion of the cathode is proportional to the $\frac{3}{2}$ power of the voltage between the plate and that part of the cathode contributing the current. In heater-type cathodes where the cathode is an equipotential surface, the total plate current for positive plate voltage is given by the equation

$$\text{Plate current} = KE_p^{3/2} \tag{57}$$

where K is a constant determined by the geometry of the tube and E_p is the anode (plate) voltage with respect to the cathode.[1]

[1] When the anode voltage is low and when very precise results are to be obtained, the voltage E_p appearing in Eq. (57) must be interpreted to mean the actual anode voltage plus a correction to take into account the contact potential existing between plate and cathode and also the effective velocity of emission of the electrons. Each of these corrections ordinarily amounts to less than 1 volt and can be neglected where the anode voltage is moderately high unless precision results are desired.

The equations giving the anode current when limited by space charge have been worked out for various structures by Langmuir, Child, Schottky, and others, and have been found to differ only in the value of the constant K appearing in Eq. (57). Derivations of these equations are to be found in numerous books on thermionic tubes, as E. L. Chaffee, "Theory of Thermionic Vacuum Tubes," Chap. IV, McGraw-Hill Book Company, Inc.; L. R. Koller, "Physics of Electron Tubes," Chap. VIII and Appendix C, McGraw-Hill Book Company, Inc.

For negative plate voltages the plate current is of course zero. In filament-type tubes the voltage drop produced in the cathode by the heating current causes different parts of the filament to have different potentials with respect to the anode, so that, while the current from each part of the filament is proportional to the $\frac{3}{2}$ power of the voltage with respect to that part, the total current is not proportional to the $\frac{3}{2}$ power of the potential of the anode. In filament-type tubes it is customary to refer the plate voltage to the negative side of the filament so that the plate is considered to be at a potential which, with respect to most of the cathode, is less than the anode voltage as measured with respect to the negative end of the filament. The result is that, when limited by space charge, the total anode current varies at a power of the anode voltage, as measured with respect to the negative side of the filament, greater than the $\frac{3}{2}$ power but which approaches the $\frac{3}{2}$ power as the voltage drop in the cathode becomes small compared with the anode voltage.[1]

28. Action of the Grid.—The flow of electrons to the plate can be controlled by placing a screenlike electrode, or grid, between the cathode and plate. This gives a three-electrode or triode tube. The grid is normally operated at a negative potential with respect to the cathode and so attracts no electrons, but the extent to which it is negative affects the electrostatic field in the vicinity of the cathode and so controls the number of electrons that pass between the grid wires and on to the plate.

[1] The way in which the total anode current varies with anode voltage when the potential drop in the filament cathode is taken into account can be readily worked out by setting up a simple differential equation based on Eq. (57). The results of such a solution give the following:

Case 1. Anode voltage less than voltage drop E_f in filament:

$$i = K\frac{2}{5E_f}E_p^{5/2} \tag{58a}$$

Case 2. Anode voltage greater than voltage drop E_f in filament:

$$\left.\begin{aligned} i &= K\frac{2}{5E_f}\left[E_p^{5/2} - (E_p - E_f)^{5/2}\right] \\ &= KE_p^{3/2}\left[1 - \frac{3}{4}\frac{E_f}{E_p} + \frac{3}{24}\left(\frac{E_f}{E_p}\right)^2 \cdots\right] \end{aligned}\right\} \tag{58b}$$

This series converges so rapidly that, when the plate voltage exceeds twice the filament drop, one can with close approximation write

$$i = KE^{3/2}\left[1 - \frac{3}{4}\frac{E_f}{E_p}\right] \tag{58c}$$

The presence of voltage drop in the cathode has the effect of causing the plate current to increase faster than the $\frac{3}{2}$ power of the voltage but at a rate that never exceeds the $\frac{5}{2}$ power. The departure from the $\frac{3}{2}$ power becomes less as the ratio E_f/E_p is reduced.

The grid may consist of any type of open-mesh structure which provides holes of ample size for the passage of electrons and which at the same time has an influence on the electrostatic field near the cathode. The most common form of grid is a coil of fine wire with widely spaced turns, as shown in Figs. 52*a* and 52*b*. The cross section of this coil is circular when a straight filament cathode or a heater cathode is used, but is usually oval when a filament in the form of a V or W is used. Several other types of grid construction in common use are also shown in Fig. 52. All these forms of grid construction function in the same way, and the choice between them is determined by such design considerations as mechanical rigidity, cost of construction, etc.

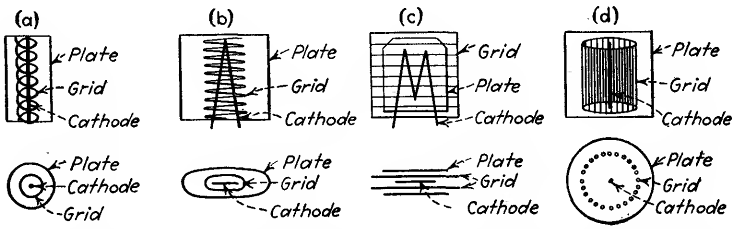


FIG. 52.—Grid, plate, and cathode structures of a number of typical tubes. It will be observed that in every case the grid is a screen-like electrode that affects the electrostatic field near the cathode while permitting electrons to flow to the plate.

The grid electrode controls the flow of electrons to the plate because the electrostatic field between the plate and cathode, and particularly near the cathode, is affected by the grid potential. This is shown in Fig. 53, which gives the electrostatic field that exists between plate and cathode for several values of grid voltage (no space charge present). When the grid is at zero potential with respect to the cathode, the positive potential of the plate produces a stray electrostatic field near the cathode, which, while somewhat weaker than would be the case with the grid removed, is still not zero because the grid is not a perfect shield. As the grid is made negative, it produces an electrostatic field between cathode and grid which opposes the stray field produced by the plate potential, and thereby weakens the electrostatic field in the vicinity of the cathode, as shown at *b* in Fig. 53. When the grid is made sufficiently negative, the stray electrostatic field produced between the cathode and grid by the positive anode is entirely neutralized by the negative grid, as shown at *c* in Fig. 53. In this last case there is no electrostatic field to draw the emitted electrons away from the cathode, and the space current will be zero.

The number of electrons that reach the anode is determined almost solely by the electrostatic field near the cathode and is affected hardly at all by the field in the rest of the interelectrode space. This is because

the electrons near the cathode are moving very slowly compared with the electrons that have traveled some distance toward the plate, with the result that the volume density of electrons in proportion to the rate of flow is large near the cathode and low in the remainder of the interelectrode space. The total space charge of the electrons in transit toward the plate is therefore made up almost solely of the electrons in the immediate vicinity of the cathode; once an electron has traveled beyond this region, it reaches the plate so quickly as to contribute to the space charge for only a brief additional time interval. The result is that the space current in a three-electrode vacuum tube is for all practical purposes determined by the electrostatic field that the combined action of the grid and plate potentials produces near the cathode.

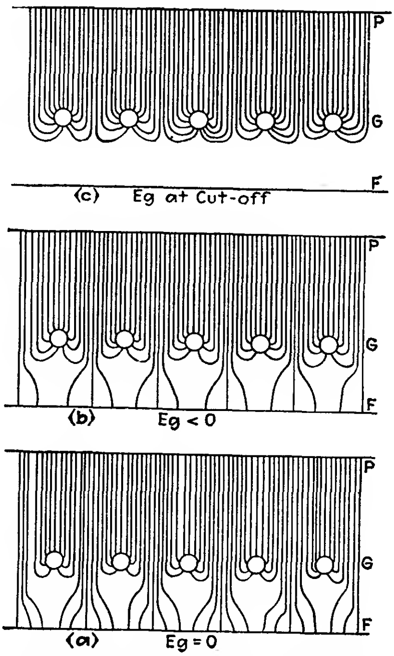


FIG. 53.—Electrostatic field produced between plate and cathode with different grid potentials, showing how the electrostatic field in the vicinity of the cathode can be controlled by the potential of the grid. These curves take into account only those fields produced by the electrode potentials, and do not include the field developed by the space charge of electrons which is superimposed upon the fields shown.

When the grid structure is symmetrical, it can be shown from the theory of electrostatics that the field at the surface of the cathode is proportional to the quantity $\left(E_g + \frac{E_p}{\mu}\right)$, where E_g and E_p are the grid and anode (plate) voltages, respectively, with respect to the cathode and where μ is a constant that is determined by the geometry of the tube and is independent of the grid and plate voltages.[1] The constant μ is known as the *amplification factor* of the tube and is a measure of the relative effectiveness

[1] The quantity $\left(E_g + \frac{E_p}{\mu}\right)$ represents the grid voltage that will produce the same electrostatic field at the surface of the cathode when the plate is at zero potential as is actually produced by the combined action of the plate and grid potentials E_p and E_g; thus it can be considered as the effective anode voltage. The combined effect of anode and grid voltages is also the same as though the grid were the only anode (*i.e.*, plate removed instead of being at the same potential as the cathode) and were at a potential of $\frac{(E_p + \mu E_g)}{1 + \mu}$.

of grid and plate voltages in producing electrostatic fields at the surface of the cathode.

Quantitative Effect of Grid Potential on Space Current.—The space current in a three-electrode tube varies with $\left(E_g + \frac{E_p}{\mu}\right)$ in exactly the same way that the space current in a two-electrode tube varies with the plate voltage, since in both cases the current flow is determined by the electrostatic field near the cathode, and this electrostatic field is in turn proportional to $\left(E_g + \frac{E_p}{\mu}\right)$ when a grid is present and to E_p when there is only a plate present. With no voltage drop in the cathode the space current is therefore proportional to $\left(E_g + \frac{E_p}{\mu}\right)^{3/2}$, and when the grid is negative all this current goes to the plate, so that for positive values of $\left(E_g + \frac{E_p}{\mu}\right)$

$$\text{Plate current} = K\left(E_g + \frac{E_p}{\mu}\right)^{3/2} \tag{59}$$

where K is a constant determined by the tube dimensions.[1] For negative values of $\left(E_g + \frac{E_p}{\mu}\right)$ the plate current is zero. It will be noted that this equation is analogous in all respects to Eq. (57) and that by interpreting $\left(E_g + \frac{E_p}{\mu}\right)$ to be the effective anode voltage they are identical. In filament-type tubes there is a voltage drop in the cathode so that different parts of the cathode are at different potentials with respect to the grid and plate. The result is that with filament-type cathodes $\left(E_g + \frac{E_p}{\mu}\right)$ must be corrected, exactly as Eq. (57) was modified to take into account the effect of the voltage drop in the filament of two-electrode tubes. It is customary to measure the grid and plate potentials with respect to the potential of the negative side of the filament, and in terms of this notation the space current of three-electrode tubes with filament-type cathodes is given by Eq. (58) provided E_f and E_p in Eq. (58) are replaced by $E_f(1 + 1/\mu)$ and $(E_g + E_p/\mu)$, respectively.

A grid maintained negative with respect to all parts of the cathode draws no electrons and so controls the plate current without consuming any power. It is this property which gives the three-electrode vacuum

[1] For highest accuracy, the parenthesis on the right-hand side of Eq. (59) must be corrected for contact potentials and velocity of emission, exactly as Eq. (57). This correction ordinarily amounts to less than 1 volt, and so is small unless the value of the parenthesis is small.

tube the ability to amplify and to generate oscillations. If the grid of the vacuum tube is allowed to go positive, it attracts large numbers of electrons and consumes a considerable amount of power, with the result that when the grid is positive the ratio of energy controlled in the plate circuit to energy consumed at the grid is small, and the amplification is reduced.[1]

29. Characteristic Curves of Triodes.—The most important characteristics of vacuum tubes with grid, plate, and cathode electrodes are the relationships between: (1) plate current and plate voltage with con-

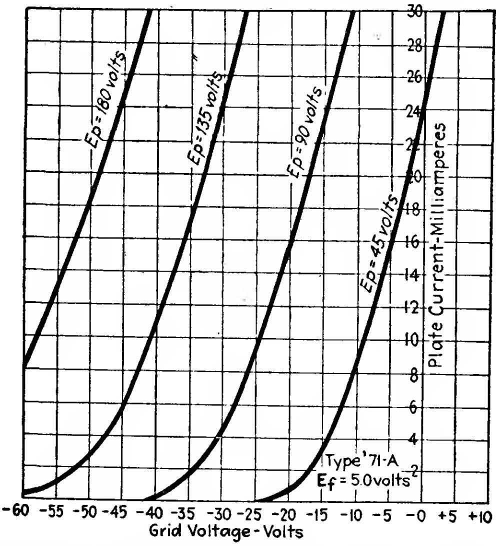


FIG. 54.—Relationship between grid voltage and plate current for several values of plate voltage in a typical three-electrode tube. Note that the only effect of changing the plate voltage is to displace the curves without changing the shape.

stant grid voltage, and (2) plate current and grid voltage with constant plate voltage. Examples of such curves are shown in Figs. 54, 55, and 56. It is to be noted that the curves showing plate current as a function of grid voltage have the same shape as those showing plate current with varying plate voltage, the only difference being in the scales involved and in the location of the curves with reference to the axes. Furthermore, the individual curves in the family showing plate current as a function of grid voltage for different values of plate voltage are of the same shape, differing only in that the curves for different plate voltages are displaced along the grid-voltage axis. The same is true of the plate

[1] The plate current appearing in Eq. (59) represents the total space current, which is the sum of the grid and plate currents. When the grid is negative, the space current is then the plate current; but when the grid is positive, the current given by Eq. (59) represents the sum of the grid and plate currents.

current-plate voltage family, where the only effect of changing the grid voltage is to displace the curve along the plate-voltage axis. It will also be noted that all these curves have the same general shape as the plate voltage-plate current curves of the two-electrode tube.

These various properties of the characteristic curves of three-electrode vacuum tubes result from the fact that the plate current is the same function of $(E_g + E_p/\mu)$ that the plate current of a two-electrode tube is of plate voltage. This means that the plate current is determined only by the value $(E_g + E_p/\mu)$ and not by the particular combination of grid and plate voltages involved. A voltage of ΔE_g added to the grid potential therefore produces the same plate-current increment as would

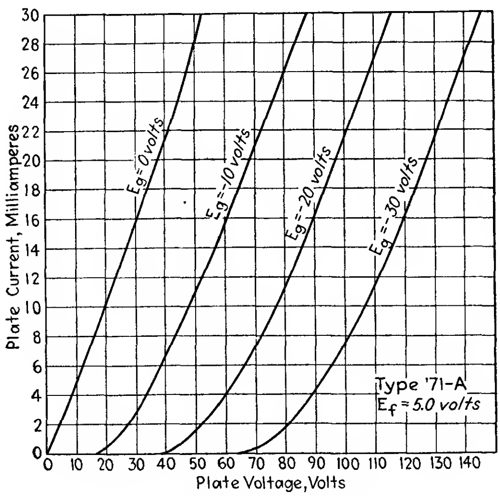


FIG. 55.—Relationship between plate voltage and plate current for several values of grid voltage for the tube used in Fig. 54. Note that the only effect of changing grid voltage (provided the grid is at least slightly negative) is to displace the curves without changing the shape; note also that these curves have the same shape as those of Fig. 54.

be produced by an increment in the plate potential of $\mu\Delta E_g$, and the effect of a grid-voltage increment ΔE_g can be neutralized by a plate-potential increment of $-\mu\Delta E_g$. In a similar manner it can be shown that changing the plate voltage ΔE_p has the same effect on the plate current as an increment $\Delta E_p/\mu$ in the grid voltage.

The range covered by Figs. 54 and 55 lies in the region where the anode current is limited by space charge. Under these circumstances the space current varies as a power of $\left(E_g + \frac{E_p}{\mu}\right)$ which ranges from $3/2$, when the effective anode voltage is large compared with $\left(1 + \frac{1}{\mu}\right)$ times the voltage drop in the filament, or when a heater-type cathode is used,

to 5/2 when the effective anode voltage is less than $\left(1 + \frac{1}{\mu}\right)$ times the voltage drop in the filament. Figure 56 shows the situation that exists when the electron emission is sufficiently low to bring in voltage saturation. It is seen that the anode current is still a function of $\left(E_g + \frac{E_p}{\mu}\right)$ exactly as in Fig. 54, but the shape of the curves is now different as a result of voltage saturation. The curves of Figs. 54 and 55 would show similar saturation effects if extended to higher values of I_p.

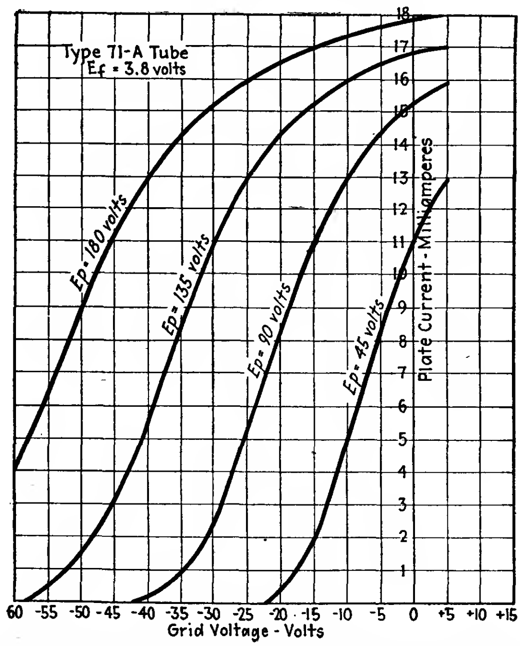


FIG. 56.—Grid-voltage plate-current curves differing from those of Fig. 54 only in that the cathode temperature has been lowered to the point where voltage saturation begins to appear at the larger plate currents, causing the tops of the curves to bend over.

The plate current of a three-element vacuum tube becomes zero when the grid is at a sufficiently negative potential barely to neutralize the stray electrostatic field produced at the surface of the cathode by the positive plate acting through the meshes of the grid. The grid potential at which this condition is realized is known as the *cutoff* grid potential and is equal to $-E_p/\mu$ volts, *i.e.*, it is the potential that makes $\left(E_g + \frac{E_p}{\mu}\right)$ equal to zero.

30. Constants of Triode Tubes.—The most important characteristics of a triode tube can be expressed in terms of three coefficients or con-

stants termed the amplification factor μ, the dynamic plate resistance R_p (generally called plate resistance), and the mutual conductance G_m. With the aid of these constants it is possible to make quantitative calculations of the tube performance under many conditions without resort to the complete characteristic curves.

Amplification Factor.—The amplification factor μ has already been defined in Sec. 28 as the ratio of the effectiveness of the grid and plate voltages in producing electrostatic fields at the cathode surface. It is determined by the geometry of the system comprising the grid, plate, and cathode electrodes, and its calculation in terms of the dimensions involved is a problem of pure electrostatics. Where the geometrical proportions can be readily introduced into an equation, as, for example, is the case when the plate and cathode are concentric cylinders and the grid is composed of a series of bars parallel to the central axis as in Fig. 52*d*, it is possible to derive formulas that will give the amplification factor with accuracy. Such ideal conditions are never completely realized in practice, however, because of supporting elements that distort the electrostatic field. The amplification factor in practical tubes is accordingly obtained by empirical formulas based upon modifications of the ideal cases.[1] The amplification factor depends primarily upon the grid structure and will be increased by anything that causes the grid to shield the cathode more completely from the plate. Thus larger grid wires or a closer spacing of the grid wires will increase the amplification factor, as does also increasing the distance between the grid and plate. The amplification factor of ordinary three-electrode tubes ranges from about 3 as the minimum to about 100 as the practical maximum, the exact value depending upon the purpose for which the tube was designed.

If the relative effects of the grid and plate voltages in producing electrostatic field at the cathode were the same for all parts of the cathode, the amplification factor μ would be absolutely independent of plate, grid, and filament voltages. In commercial tubes various mechanical requirements, such as the necessity of supporting wires and the inevitable imperfections in construction, result in dissymmetry that causes different parts of the tube to have somewhat different amplification factors. The over-all amplification factor of such a combination will vary with plate, grid, and filament voltages and will tend to become lower as cut-off

[1] A comprehensive treatment of amplification-factor computations is given by Yuziro Kusunose, Calculation of Characteristics and the Design of Triodes, *Proc. I.R.E.*, vol. 17, p. 1706, October, 1929. For additional information the reader is referred to the following: R. W. King, Calculation of the Constants of Three-electrode Thermionic Vacuum Tube, *Phys. Rev.*, vol. 15, p. 256, April, 1920; John M. Miller, The Dependence of the Amplification Constant and Internal Plate Circuit Resistance of Three-electrode Vacuum Tubes upon the Structural Dimensions, *Proc. I.R.E.*, vol. 8, p. 64, February, 1920.

is approached because, as the grid becomes more negative, those parts having the highest value of μ will reach cut-off first, leaving only the low μ parts of the tube contributing to the space current. The way in which the amplification factor varies over the characteristic curves of a typical tube is shown in Fig. 57. It is seen that over the main part of the characteristic, which also represents the usual operating range, the amplification factor does not vary by more than 10 to 15 per cent, but that at very low plate currents the variation in the amplification factor is greater. The results shown in this figure are typical of a large number of tubes that have been investigated.[1]

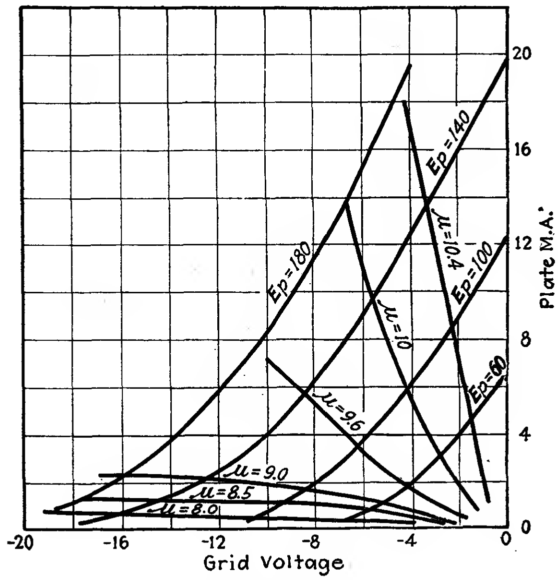


FIG. 57.—Curves showing variations in amplification factor in a tube having no voltage drop in the cathode (heater-type cathode). The amplification factor is relatively constant over most of the region of negative grid potential but tends to become less as the plate current becomes less, particularly when the grid is highly negative at the same time.

In the practical case where the amplification factor is not a pure geometrical constant, it is defined as the relative effectiveness of grid and plate voltages in controlling the plate current, and so is expressed by the following mathematical relation:

$$\text{Amplification factor} = \mu = \frac{\partial I_p/\partial E_g}{\partial I_p/\partial E_p} = -\left.\frac{dE_p}{dE_g}\right|_{I_p \text{ constant}} \tag{60}$$

[1] A more complete discussion of the causes of the dependence of the amplification factor upon electrode voltages is to be found in F. E. Terman and A. L. Cook, Note on Variations in the Amplification Factor of Triodes, *Proc. I.R.E.*, vol. 18, p. 1044, June, 1930.

Dynamic Plate Resistance.—The dynamic plate resistance of a vacuum tube represents the resistance that the plate circuit offers to a small increment of plate voltage. Thus, when an increment of plate voltage ΔE_p produces an increment in the plate current of Δi_p, the dynamic plate resistance is given by the relation

$$\text{Dynamic plate resistance} = R_p = \frac{\Delta E_p}{\Delta I_p} = \frac{\partial E_p}{\partial I_p} \tag{61}$$

The plate resistance is therefore the reciprocal of the slope of the plate current-plate voltage characteristic shown in Fig. 55 and depends upon the grid and plate voltages at the operating point under consideration. It is important to remember that the plate resistance is determined by the slope of the plate voltage-current curve, being lowest at points where the slope is greatest, *and is not equal to the ratio of total plate voltage to total plate current.*

In any particular tube the dynamic plate resistance depends primarily upon the plate current and only to a small extent upon the combination of grid and plate voltages used to produce this current. Furthermore, the dynamic plate resistance becomes progressively lower as the plate current is increased. This behavior is a result of the fact that the different plate voltage-plate current curves in Fig. 55 differ primarily in being displaced along the plate-voltage axis and in that the slope of each curve increases as the plate current becomes greater.[1] The result is that the plate resistance R_p tends to be the same for all curves with a given current and becomes less as the current is increased. This ideal is not entirely realized in practical tubes as a result of the same irregularities that prevent the amplification factor from being constant, so that in commercial tubes the plate resistance is to a certain extent affected by the combination of grid and plate voltages involved, as well as by the value of plate current. The characteristics of an actual commercial tube are shown in Fig. 58 for a representative case.

[1] This can also be demonstrated by use of Eq. (59). Thus

$$R_p = \frac{\partial E_p}{\partial I_p} = \frac{1}{\partial I_p/\partial E_p} = \frac{1}{\frac{3K}{2\mu}\left(E_g + \frac{E_p}{\mu}\right)^{1/2}}$$

Expressing $\left(E_g + \frac{E_p}{\mu}\right)^{1/2}$ in terms of $\left(\frac{I_p}{K}\right)$ from Eq. (59) gives

$$R_p = \frac{2\mu}{3K^{2/3}}\left(\frac{1}{I_p^{1/3}}\right)$$

Hence the plate resistance is approximately inversely proportional to the cube root of the plate current.

The plate resistance of a triode tube depends upon the dimensions and relative positions of the cathode, plate, and grid. It becomes less as the effective cathode area is increased, *i.e.*, as the cathode is made longer or of larger diameter and as the distance between the cathode and the other electrodes is reduced. The dynamic plate resistance can be computed with accuracy by theoretical formulas only for the same simple and symmetrical geometrical conditions for which the amplification factor can be determined by exact methods. In commercial tubes it is necessary to use empirical modifications of these formulas to take into account the effect of dissymmetries in the geometry. This is usually done by developing empirical rules for defining an effective cathode area which, when substituted in the theoretical formula, will give results in agreement with those actually observed.[1] Since the power required to heat the cathode depends upon the surface area of the cathode, it may be said in a general way that the plate resistance is dependent upon the cathode heating power and is lowered by employing a larger cathode.

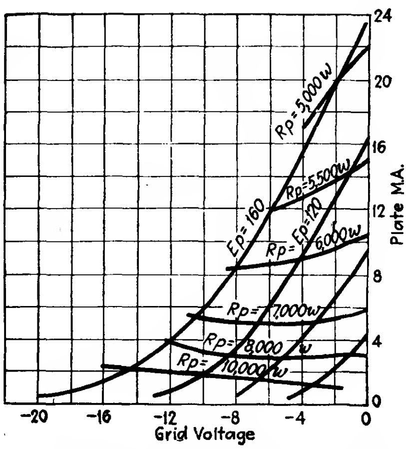


FIG. 58.—Curves showing variations in the plate resistance of a tube having no voltage drop in the cathode. The resistance is approximately the same for a given plate current irrespective of the combination of plate and grid voltages used to produce this current.

The plate resistance of tubes differing only in grid structure decreases as the amplification factor is lowered. This is because the electrostatic field produced near the cathode by a given plate voltage is inversely proportional to the amplification factor. As a consequence a given increment of plate voltage produces an increment in the electrostatic field at the surface of the cathode (and consequently an increment in the plate current) which varies inversely with the amplification factor. Tubes having no voltage drop in the cathode and differing only in the grid structure will have a plate resistance that is approximately directly proportional to the amplification factor of the tube when compared at operating points having the same plate current.

Mutual Conductance (or Transconductance).—The mutual conductance G_m (or, as it is often called, the transconductance S_m) is defined as the rate of change of plate current with respect to a change in grid voltage. Thus, if the grid voltage is changed by ΔE_g, the resulting plate-current

[1] For further information regarding such calculations see Yuziro Kusunose, *loc. cit.;* R. W. King, *loc. cit.*

change ΔI_p is related to the mutual conductance by the equation

$$\Delta i_p = \Delta E_g G_m$$
$$G_m = \frac{\partial I_p}{\partial E_g} = \left.\frac{dI_p}{dE_g}\right|_{E_p \text{ constant}} \tag{62a}$$

By combination of Eqs. (60) and (61) it is also found that the mutual conductance is the ratio of amplification factor to plate resistance. That is

$$G_m = \frac{\mu}{R_p} = \frac{\partial I_p}{\partial E_g} \tag{62b}$$

The mutual conductance is a rough indication of the design merit of a tube. This is because a low plate resistance and a high amplification factor are desired, and the mutual conductance measures the extent to which this feature is attained. Tubes of equal design merit but with slightly different values of amplification factor will have substantially the same value of mutual conductance under normal operating conditions. When tubes with widely different values of μ are compared, the tendency is for the mutual conductance to be less for the high-amplification-factor tubes, and this effect is especially pronounced when there is a voltage drop in the cathode.

In a particular tube the mutual conductance depends primarily upon the plate current, and to only a small extent upon the combination of grid and plate voltages used to produce this current. The mutual conductance also increases as the plate current is increased. This follows from the fact that $G_m = \mu/R_p$, so that, as the amplification factor is substantially constant, the mutual conductance varies inversely as R_p.

31. Pentodes.—A pentode is a five-electrode tube consisting of cathode, plate, and three grids which are concentrically arranged between cathode and plate as in Fig. 59. The inner grid is called the control grid and corresponds to the grid of a triode tube. The next grid is termed the screen, or screen grid, while the outer grid is called the suppressor. In normal operation the control-grid is maintained negative at all times with respect to the cathode. The screen grid is operated at a fixed positive potential while the suppressor is normally connected directly to the cathode. The plate is operated at a positive potential.

The main differences between the pentode and triode tubes are the two extra grids of the pentode and the fact that there are two positive electrodes. These differences modify the relationship that exists between plate current, control-grid voltage, and plate voltage in a way that is very desirable for many purposes. The two extra grids also provide electrostatic shielding, which eliminates substantially all capacity coupling between the plate and the control grid. In radio-frequency

amplifiers this shielding is extremely desirable, and, in order that it may be as nearly perfect as possible, pentode tubes for radio-frequency work are constructed as shown in Fig. 59. Here the control-grid lead is brought out through the top of the tube and there is in addition a

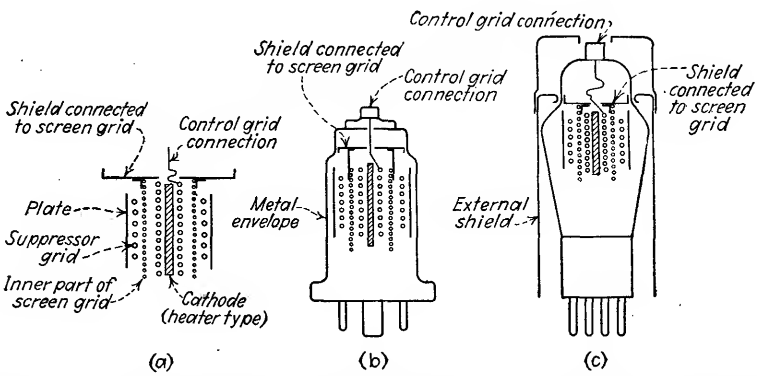


FIG. 59.—Schematic diagrams of typical pentode tubes, showing the electrode arrangement and how the control grid is completely shielded from the plate in tubes intended for use at high frequencies. Audio-frequency power pentodes are similar, except that the shield is omitted.

metallic shield connected to the top of the screen for the purpose of providing additional electrostatic shielding between the control-grid lead and the plate. In metal tubes this shield cooperates with the metal envelope to give substantially complete shielding between control grid and plate, as shown in Fig. 59*b*, while in glass tubes an external metal shield is arranged to accomplish substantially the same result, as illustrated in Fig. 59*c*. In audio-frequency power pentodes this shield is omitted.

Factors Controlling Voltage and Current Relations.—The nature of the voltage and current relations existing in a pentode tube can be worked out by studying the electrostatic fields that exist between the electrodes. These fields are shown in Fig. 60 for a typical case. The strength of the field at the surface of the cathode determines the number of electrons that are drawn away from the space charge about the cathode, exactly as in the case of the triode tube. This field is determined by the screen-grid and control-grid potentials and by the geometry of the tube. It is not affected by the plate potential because the screen and suppressor grids effectively

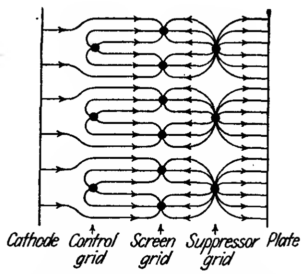


FIG. 60.—Schematic diagram illustrating the nature of the electrostatic fields existing in a pentode tube. The arrows indicate the direction in which the electrons are accelerated.

shield the plate, with the result that the number of electrons drawn away from the space charge is substantially independent of the plate voltage. The electrons drawn from the space charge pass between the control-grid wires and are accelerated to a high velocity as the screen grid is approached. At this high velocity the electrons travel in substantially straight lines, so very few except those which happen to be going directly toward the screen-grid wires are intercepted by the screen. The remaining electrons then pass through the screen grid and travel on toward the suppressor. As the suppressor is approached, the electrons slow down because of the retarding field between suppressor and screen; but if the plate is reasonably positive, they will pass on through the spaces between the suppressor grid wires and reach the plate without stopping. This is because the positive plate is able to produce appreciable electrostatic field on the screen side of the suppressor grid so that, when the electrons reach the influence of this field, they are drawn through the spaces between the suppressor grid wires and are accelerated toward the plate. The only electrons actually reaching the suppressor are therefore an occasional few which were emitted from the cathode with unusually high velocity of emission and which happened by chance to be directed exactly toward the middle of the suppressor wires.

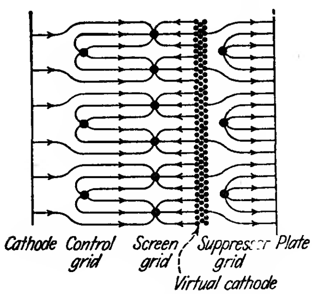


FIG. 61.—Schematic diagram of electrostatic fields existing in a pentode tube when a virtual cathode is formed in front of the suppressor. The arrows indicate the direction in which the electrons are accelerated.

When the plate voltage is very low, and particularly when the total space current is large at the same time, the above behavior is modified because the plate is then not capable of producing sufficient electrostatic field on the screen side of the suppressor to draw off the electrons as, rapidly as they arrive. The result is that a space charge forms in front of the suppressor, as shown in Fig. 61. This space charge is called a *virtual cathode*, and, when the accumulation of electrons is sufficient, it acts in all respects as does the space charge surrounding the actual cathode. In particular, the virtual cathode in conjunction with the plate and suppressor grid forms the equivalent of a triode tube in which the suppressor is the grid. The number of electrons that reach the plate under such conditions is a function of the suppressor and plate voltages, and tends to be independent of the control-grid and screen-grid potentials provided the virtual cathode has a complete space charge. The excess electrons which the plate is not capable of attracting from the virtual cathode return toward the emitting cathode and are collected by the cathode and screen electrodes.

It follows from the above discussion that, when no virtual cathode is formed, the total space current in the case of an equipotential (heater) cathode is given by the equation[1]

$$\text{Total space current} = I_p + I_{sg} = K\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)^{3/2} \tag{63}$$

where

I_p and I_{sg} = plate and screen currents, respectively
K and μ_{sg} = constants determined by the tube construction
E_g and E_{sg} = control-grid and screen-grid potentials, respectively.

The plate voltage has practically no effect and so does not appear in the relation.

Equation (63) results from the fact that the space current is proportional to the $3/2$ power of the electrostatic field at the cathode surface. This field is determined only by the control- and screen-grid potentials,[2] with the screen grid serving the same function as the plate of the triode in Eq. (59). The quantity μ_{sg} is analogous to the amplification factor of the triode, and with perfect symmetry is a constant determined solely by the geometry of the tube. The numerical value of μ_{sg} is a measure of the extent to which the control grid is able to shield the cathode from the screen-grid potential, and is greater the finer the mesh of the control-grid structure. It is apparent from Eq. (63) that the total space current can be controlled by means of the control grid and that this current becomes zero, or is cut off, when the control-grid bias is

$$\text{Cutoff bias} = \frac{-E_{sg}}{\mu_{sg}} \tag{64}$$

Because of this relation one can think of μ_{sg} as the "cutoff" amplification factor.

The total space current given by Eq. (63) is divided between the positive electrodes, *i.e.*, between the screen and the plate. With plate voltages that are sufficient to prevent the formation of a virtual cathode in front of the suppressor, the ratio of plate to screen currents is very nearly equal to the ratio of the area of the spaces between the wires of the screen-grid structures to the projected area of the wires themselves, and is to a first approximation independent of the plate and screen potentials.[3]

[1] For highest accuracy the quantity inside the brackets on the right-hand side of the equation must be corrected to take into account contact potential and velocity of emission, exactly as discussed in connection with Eq. (59) for the case of triodes.

[2] This assumes that the suppressor is connected to the cathode. If the suppressor grid is not connected to the cathode, its potential has an effect which, though very small, is still detectable.

[3] The only effects that can alter the division of space current between plate and screen, when the plate voltage is adequate to prevent the formation of a virtual cathode

Characteristic Curves of Pentodes.—The actual voltage and current relations of pentode tubes can be shown in characteristic curves, of which those shown in Figs. 62 to 64 are typical. These are consistent in every detail with the theoretical explanation given above. The total space current tends to be independent of plate voltage and to vary with screen- and control-grid potentials in exactly the same way as the plate current of a triode tube. The division of this total space current between plate and screen is also seen to be to a first approximation independent of plate, control-grid, and screen potentials provided the plate potential is not too low. As a consequence the plate current and also the screen-grid current vary with control-grid and screen-grid voltages in the same way as does the total space current (except when the plate potential is low). At plate

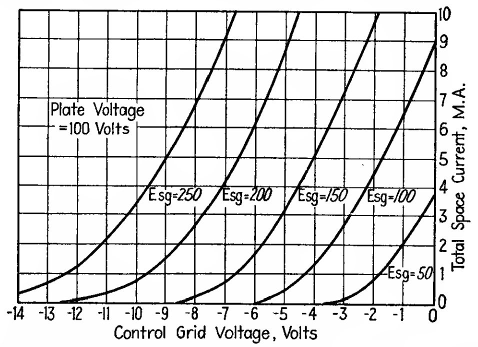


FIG. 62.—Curves showing total space current of a pentode as a function of control-grid potential for various screen potentials. Note the similarity of these curves to those of Fig. 54 for triodes.

voltages so low that an effective virtual cathode is formed in front of the suppressor, the plate current tends to be independent of control-grid and screen-grid potentials and to be determined by the plate voltage, as theory indicates should be the case. The screen and cathode then divide between them the part of the total space current that does not go to the plate, so that the screen current increases and the total net space current decreases as the plate voltage is reduced.

---

in the front of the suppressor grid, are: (1) the fact that an increased plate potential may divert a few of the small number of electrons that would otherwise go to the suppressor; (2) the fact that an increased plate potential will modify the electrostatic fields existing at the screen grid, and hence may divert to the plate a few electrons that would otherwise be intercepted by the screen; and (3) the fact that the attraction of the screen-grid wires does deflect slightly the paths of the approaching electrons, causing the screen potential to have a small influence on the number of electrons that are intercepted. All these effects are relatively small, however, so that to a first approximation the division of current between plate and screen is roughly independent of the control-grid, screen, and plate potentials provided the plate potential is not too low.

In some circumstances the suppressor grid is biased negatively and is used to control the plate current. When this negative bias is sufficient to

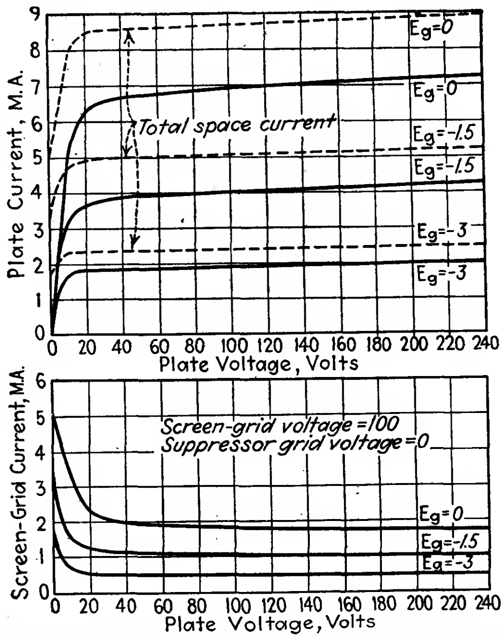


FIG. 63.—Curves showing plate and screen currents and total space current of a pentode as a function of plate voltage for various control-grid potentials. Note that when the plate potential is not too low the plate voltage has relatively little effect on the currents.

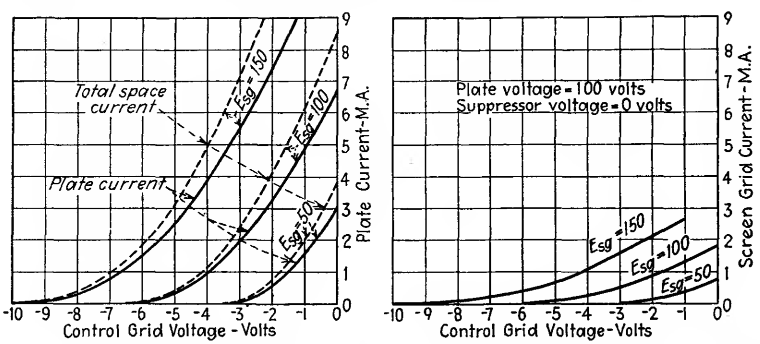


FIG. 64.—Curves showing plate and screen currents and total space current of a pentode as a function of control-grid voltage for various screen potentials. Note that these curves are of the same general character as those for the total space current of a pentode (Fig. 62) and those of plate current of a triode (Fig. 54).

permit the formation of an effective virtual cathode in front of the suppressor, the plate current behaves in much the same way as it would in a triode tube in which the suppressor grid was the control grid and the

virtual cathode the actual cathode. This is clearly apparent in Fig. 65. In order to have an effective virtual cathode, it is necessary that the total space current be appreciably greater than the number of electrons that the plate is able to draw from the virtual cathode. When the virtual cathode is not able to provide a complete space charge, saturation effects appear which correspond to those observed in a triode with insufficient cathode emission, as is evident by a comparison of Figs. 65 and 56. For any given total space current this saturation effect first appears at a

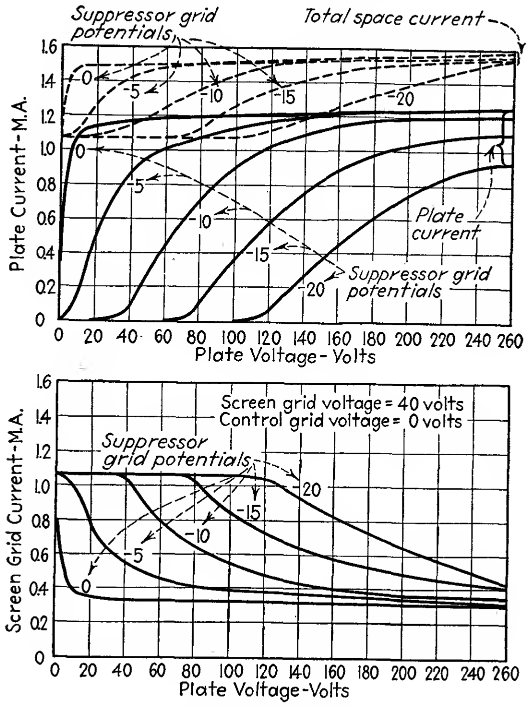


FIG. 65.—Curves showing plate and screen currents and total space current of a pentode as a function of plate voltage for various values of suppressor-grid potential. It will be noted that for low plate voltages where a virtual cathode forms, the curves of plate current are somewhat like those of a triode as shown in Fig. 56.

value of plate current which is proportional to the total space current, but which is roughly independent of the combination of plate and suppressor voltages used to draw the plate current. Hence the tendency toward the formation of a virtual cathode increases as the total space current is increased and as the plate current becomes smaller. Examination of Fig. 65 shows that, as the suppressor potential is varied, the total space current also varies. At plate voltages high enough to prevent the formation of a virtual cathode, this effect is small and arises from the fact that the suppressor potential has some effect on the field near the cathode

in spite of the shielding action of the screen grid. At plate voltages low enough to allow a virtual cathode to form, the total space current decreases rapidly as the suppressor is made more negative. This is because many of the excess electrons which the virtual cathode turns back toward the screen pass on between the control-grid wires and are collected by the emitting cathode, and so are not counted in the total space current.

One may wonder why the suppressor grid is necessary in the pentode tube. Actually the suppressor has a very important function because, with the electrostatic fields as shown in Fig. 60, secondary electrons produced at the plate and screen grid by the impact of the electrons arriving from the cathode will be immediately drawn back to the electrode produc-

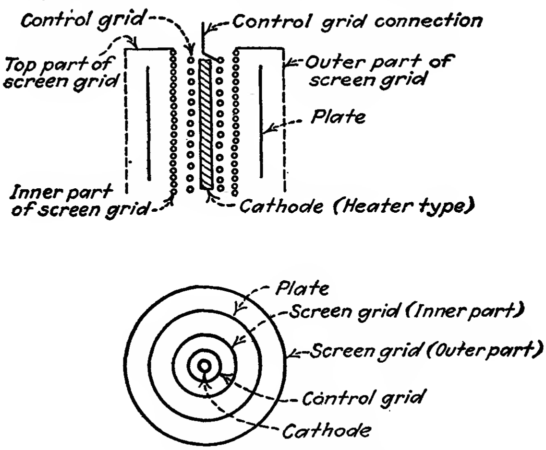


FIG. 66.—Section taken through a screen-grid tube to show the electrode arrangement.

ing them. If the suppressor grid is omitted, however, there will be strong electrostatic fields existing between the plate and screen electrodes, and whichever of these electrodes is the most positive will attract the secondary electrons produced at the less positive electrode. This introduces undesirable characteristics and is the reason that the screen-grid tube, which can be thought of as a pentode with the suppressor grid omitted, has been displaced in nearly all applications by the pentode tube.

32. Screen-grid Tubes.—A screen-grid tube is essentially a three-electrode tube to which there has been added a second grid located between the plate and the first grid. This extra electrode is called the screen grid, and it serves as an electrostatic shield that eliminates substantially all capacitive coupling between the inner grid and plate. In order that this shielding may be as complete as possible, small screen-grid tubes are commonly constructed as shown in Fig. 66, with the control-grid lead coming out through the top. The extension of the screen grid that practically surrounds the plate assists in reducing the direct capacity between

control grid and plate to the lowest possible value. The screen is normally operated at a fixed positive potential somewhat less than the plate voltage, but most of the electrons that the screen attracts pass through its meshes and go to the plate, so that, while the screen grid serves as an almost perfect electrostatic shield, it intercepts only a relatively small proportion of the total electron flow.

The voltage and current relations existing in a screen-grid tube can be derived by considering the electrostatic fields that exist in the tube and the effect that these have upon the behavior of the primary and secondary electrons. The electrostatic fields for two typical cases are shown schematically in Fig. 67. The number of electrons drawn from the space charge surrounding the cathode is determined by the electrostatic field in the immediate vicinity of the cathode, and this in turn is determined by

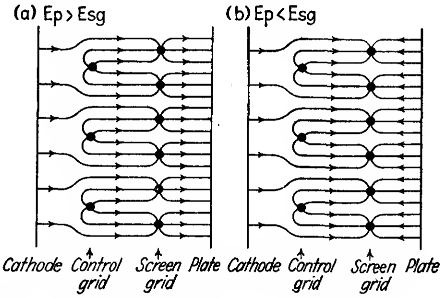


FIG. 67.—Schematic diagram illustrating the nature of the electrostatic fields existing in a screen-grid tube. The arrows indicate the direction in which the electrons are accelerated.

the control-grid and screen-grid potentials exactly as in the case of the pentode tube. The potential of the plate has negligible effect upon this field because of the effective electrostatic shielding exerted by the combined action of screen and control grids. The electrons drawn from the space charge pass between the meshes of the control grid and are accelerated to a high velocity on approaching the screen. Most of these electrons also pass between the meshes of the screen grid and go on to the plate. The total number of electrons drawn from the space charge surrounding the cathode, and the division of these electrons between the plate and screen-grid electrodes, are determined by exactly the same factors that control the total space current and its division in pentode tubes.

The actual plate and screen-grid currents of the screen-grid tube are not, however, determined solely by the number of electrons that these electrodes receive directly from the cathode. This is because the electrons upon arriving at the screen grid and plate produce secondary electrons by impact (secondary emission). These secondary electrons are then attracted to the electrode having the highest positive potential, with the

result that there is an electron flow between screen grid and plate which is superimposed upon the flow of primary electrons to these electrodes from the cathode. The more positive electrode consequently receives a larger current and the less positive electrode a smaller current than would be the case if there were no secondary emission. Thus, when the plate potential is appreciably less than the screen potential, as in Fig. 67*b*, the screen attracts all the secondary electrons produced at the plate. The net plate current is then the number of primary electrons received minus the number of secondaries lost, and it will be negative when each arriving primary electron produces on the average more than one secondary electron. Similarly, when the plate is more positive than the screen grid,

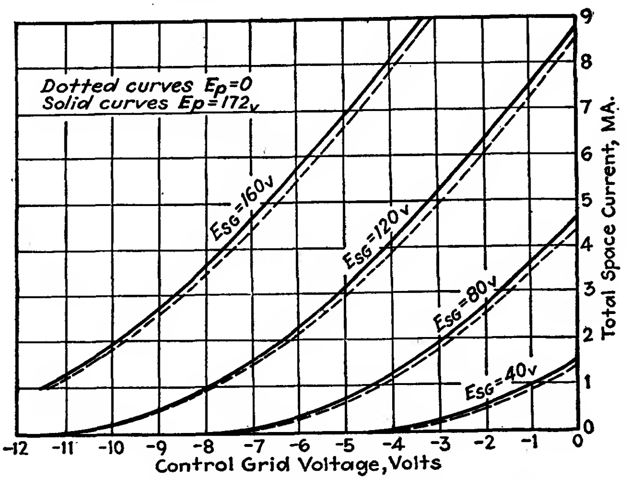


FIG. 68.—Typical curves showing the effect of electrode voltages on the total space current ($i_p + i_{sg}$) of a screen-grid tube. This total current is relatively independent of the plate potential and varies with control- and screen-grid potentials in exactly the same way as the plate current of a triode varies with grid and plate voltages, respectively.

as in Fig. 67*a*, the plate attracts secondary electrons from the screen, but this effect is not so pronounced since most of the secondaries at the screen are produced on the side away from the plate. The screen, therefore, largely shields these secondaries from the attraction of the plate which enables the screen to recapture most of them.

The number of secondary electrons produced at an electrode is proportional to the number of arriving primary electrons, increases as the voltage of the electrode becomes greater, and is particularly sensitive to the surface conditions. Secondary emission commonly becomes appreciable at potentials of 25 to 75 volts, and at these voltages it is not unusual for each primary electron to produce one or two secondary electrons. With surfaces treated in such a way as to enhance secondary emission, as many as ten secondary electrons may be produced for each primary elec-

tron, while surfaces prepared to resist secondary emission will on the average have only one secondary electron for perhaps five or ten primary electrons.

Characteristic Curves of Screen-grid Tubes.—The above factors cause the screen-grid tube to have characteristic curves such as shown in Figs. 68 to 71. The curves of total space current in Fig. 68 are similar to the corresponding curves for the pentode tube shown in Fig. 62 and are similar to the characteristic curves of a triode with the screen grid serving the same function as the triode plate. For an equi-potential cathode one hence has[1]

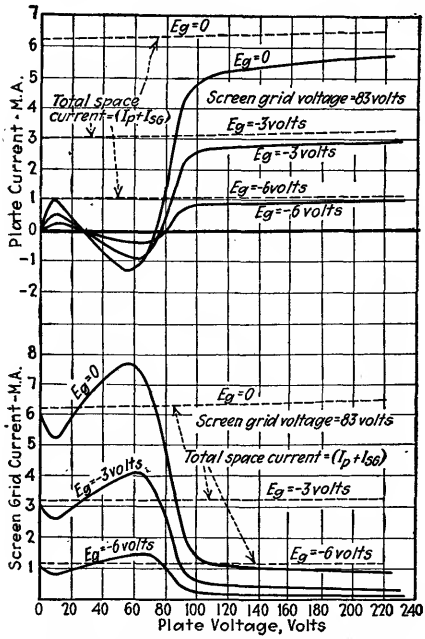


FIG. 69.—Variation of plate and screen-grid currents and of total space current, with plate voltage (screen-grid voltage constant). It will be noted that changing the control-grid voltage alters the magnitude of the curves without changing their shape (*i.e.*, the control-grid potential affects the total space current but does not alter its division between the plate and screen grid).

Total space current =

$$I_p + I_{sg} = K\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)^{3/2} \qquad (65)$$

The notation is identical with that used in Eq. (63) for pentodes. The plate potential has negligible effect on the total space current because of the effective shielding provided by the screen grid.[2]

The division of the total space current between the plate and screen electrodes depends upon the relative and absolute voltages of those electrodes as shown in Figs. 69, 70, and 71. When the plate is appreciably more positive than the screen electrode, the plate retains all the

[1] For greatest accuracy, corrections for contact potential and velocity of emission must be included in the parenthesis on the right-hand side of Eq. (65), as was the case in Eqs. (59) and (63).

[2] To take into account the effect of the plate potential, Eq. (65) can be written in more exact form as follows:

$$I_p + I_{sg} = K\left(E_g + \frac{E_{sg}}{\mu_{sg}} + \frac{E_p}{\mu_p}\right)^{3/2}$$

where μ_p is a constant that is a measure of the relative effectiveness of the control grid and the plate in producing electrostatic fields at the surface of the cathode. This constant μ_p is consequently a measure of the shielding, and will be so large as to make the term E_p/μ_p negligible if the shielding is of the order of magnitude obtained with ordinary screen-grid tubes.

primary electrons that it receives and in addition receives a small number of secondary electrons from the screen, with the result that the plate current is then very nearly equal to the total space current and will be substantially independent of plate voltage.[1] When the plate voltage is less than the screen potential and is still not extremely low, the plate current decreases as the plate voltage increases, as in Fig. 69. This represents a *negative resistance* characteristic, and the tube, when used in this way as a negative resistance device, is termed a *dynatron*. The

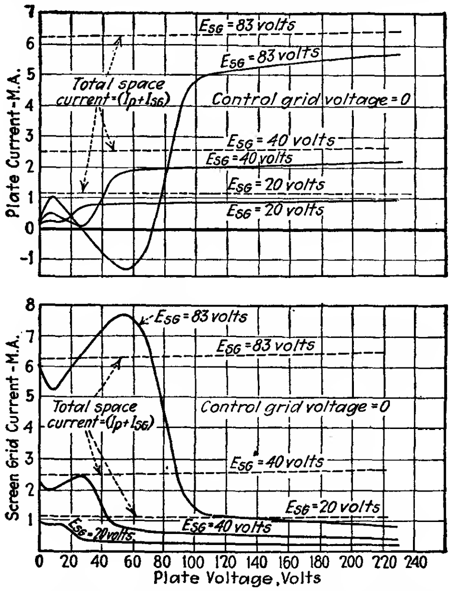


FIG. 70.—Variation of plate and screen-grid currents and total space current with plate voltage for several values of screen-grid voltage (control-grid voltage constant). The total space current is nearly independent of plate voltage, but the way in which this current divides between screen grid and plate depends upon the potentials of these electrodes, the one which is the most positive receiving the major fraction of the current.

negative resistance is a result of the fact that, while the number of primary electrons that the plate receives is independent of the plate voltage, the number of secondary electrons produced at the plate increases with increased plate voltage; and with the screen more positive than the plate all these secondary electrons flow to the screen. Hence the plate current decreases with increasing plate voltage, and will reverse in polarity if on

[1] When the plate voltage is appreciably greater than the screen voltage, the plate current will vary slightly with plate voltage because increasing the plate voltage (1) increases the total electron flow very slightly as a result of imperfect shielding by the screen grid, (2) increases slightly the fraction of the total space current that the plate receives as a result of the electrostatic fields which the plate produces at the screen, and (3) increases that fraction of the secondary electrons produced at the screen grid which the plate is able to attract.

the average each primary electron produces more than one secondary electron. The transition between this region of decreasing plate current with increasing plate voltage and the region where the plate current equals substantially the total space current and is independent of the

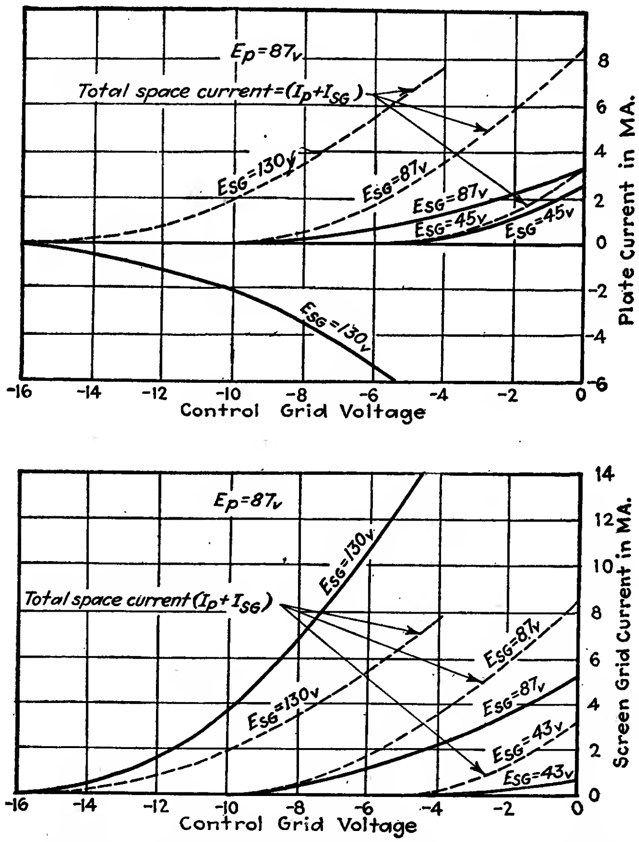


FIG. 71.—Variation of plate and screen-grid currents and total space current with control-grid voltage for several values of screen-grid voltage. These curves bring out how, for a given plate and screen-grid voltage, the absolute magnitude of the current received by each of these electrodes varies with control-grid voltage in much the same way as the plate current of a triode varies with grid voltage.

plate voltage occurs when the screen and plate are at approximately the same potential.

At very low plate voltages the number of secondary electrons produced at the plate becomes very small and at the same time there is a tendency for a space charge (virtual cathode) to form in front of the plate that turns back some of the arriving electrons to the screen. The plate current then depends on the plate voltage and is much less than the total space current.

The screen grid gets whatever current does not go to the plate, since the arithmetic sum of plate and screen currents must equal the total space current. As a result the screen current varies with plate voltage in a manner that is the exact inverse of the way in which the plate current varies with plate voltage.

Increasing the screen voltage increases the total space current and raises to a higher value the plate voltage at which the plate current first becomes substantially independent of plate voltage. At the same time the region where the plate current decreases with increasing plate voltage is lengthened. This is clearly evident in Fig. 70.

The control-grid potential alters the magnitude of plate and screen current curves but does not affect their general shape. This is because

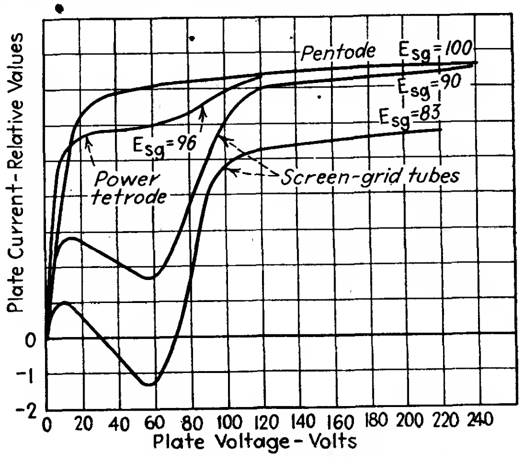


FIG. 72.—Characteristics of screen-grid and pentode tubes showing the effects of secondary emission. The power tetrode is a screen-grid tube in which the plate has a ribbed structure and is treated to reduce the production of secondary electrons.

the control-grid potential affects the number of primary electrons drawn from the space charge around the cathode but does not appreciably affect the factors that control the division of current between the screen and plate. The result is that the plate and screen currents, as well as the total space current $(I_p + I_{sg})$, vary with control-grid voltage in much the same way as does the plate current of a triode. This is illustrated in Fig. 71 and is true even when the absolute magnitude of the current to one electrode is negative.

Miscellaneous Considerations.—A comparison of screen-grid and pentode tubes shows that the differences between them arise primarily from the fact that the suppressor grid of the pentode prevents secondary electrons from being interchanged between plate and screen. This eliminates the fold that appears in the plate-current curves of Fig. 69, makes the plate current somewhat more independent of plate voltage, and

also makes possible slightly better shielding between plate and control grid. The result is that for most purposes the pentode tube is superior to the screen-grid tube.

The characteristics of screen-grid tubes in the region where the plate voltage is less than the screen voltage are very sensitive to the surface conditions at the plate electrode. These conditions vary greatly from tube to tube, and also often change during the life of a single tube. It is possible to prevent almost all loss of secondary electrons at the plate by treating the plate surface in such a way as to reduce the tendency toward secondary emission, and by providing the plate with a ribbed structure that serves to shield the plate surface from the screen-grid potential and so enables the plate to retain a considerable fraction of the secondary electrons produced at its surface. The characteristic curve of a tube of this type is shown in Fig. 72 (power tetrode curve), together with the corresponding characteristics of two other screen-grid tubes having different amounts of secondary emission, and the characteristics of a pentode tube.

33. Coefficients of Screen-grid and Pentode Tubes.—The important coefficients of screen-grid and pentode tubes are the mutual conductance (or transconductance), plate resistance, and two amplification factors. The first amplification factor of importance is the constant μ_{sg} which appears in Eqs. (63) and (65) and which represents the relative effectiveness of the control and screen grids in producing electrostatic fields at the surface of the cathode. A definition of this constant in terms of the notation of Eqs. (63) and (65) is

$$\mu_{sg} = -\frac{dE_{sg}}{dE_g}\bigg|_{I_p + I_{sg} \text{ constant}} \tag{66}$$

The amplification factor μ_{sg} is analogous to the amplification factor μ of triode tubes, and can be termed the *cut-off amplification factor* since it determines the screen- and control-grid potentials giving cut-off. It can be calculated in the same manner as the amplification factor of a triode by assuming that the screen corresponds to the plate electrode of the triode. If the tube has perfect mechanical symmetry, the amplification factor μ_{sg} is determined only by the geometry of the tube and is completely independent of electrode voltages. However, the dissymmetry introduced by supporting wires causes different parts of the control grid to have different values of μ_{sg}, so that the resulting value of this constant for the entire tube is a sort of average that decreases in numerical value as cut-off is approached, exactly as is the case with the amplification factor of triodes. The numerical value of μ_{sg} in receiving screen-grid and pentode tubes is commonly in the range 6 to 15, with slightly lower values sometimes being employed in screen-grid transmitting tubes.

The second amplification factor of importance in screen-grid and pentode tubes represents the relative effectiveness of control-grid and plate potentials in varying the plate current. This constant is commonly termed the amplification factor, is designated by the symbol μ, and is defined quantitatively by the equation

$$\mu = -\frac{dE_p}{dE_g}\bigg|_{I_p \text{ constant}} \tag{67}$$

where E_p and E_g are plate and control-grid potentials, respectively, and I_p is the plate current. Over that part of the characteristic curves of pentodes and screen-grid tubes where the plate current is substantially independent of plate voltage, the amplification factor μ will be extremely high because the plate potential then has very little effect upon the plate current, whereas the control grid will have a very large effect. In radio-frequency pentodes the value of μ commonly encountered exceeds 1000 except at very low plate voltages, while, even in audio-frequency pentodes where the shielding of the plate electrode is far from perfect, the value of μ is commonly several hundred. Values for screen-grid tubes are normally slightly less than with radio-frequency pentodes when the plate potential is appreciably higher than the screen voltage. When the plate voltage is very low in pentodes, or is equal to or less than the screen voltage in screen-grid tubes, the amplification factor μ becomes low and will vary greatly with electrode voltages. The amplification factor μ is not a geometrical constant as is μ_{sg} ,but rather depends upon the electrode voltages.

Plate Resistance.—The plate resistance of screen-grid and pentode tubes represents the resistance that the plate circuit offers to an increment in plate potential, exactly as in the case of triodes. Thus it is defined by the equation

$$\text{Plate resistance} = R_p = \frac{\partial E_p}{\partial I_p} \tag{68}$$

The plate resistance is the reciprocal of the slope of the plate-voltage–plate-current curve, and hence is very high when the plate current is nearly independent of plate voltage. In radio-frequency pentodes the plate resistance exceeds 1 megohm except at very low plate voltages, and, even in audio-frequency pentodes where the effect of the plate potential is not so completely shielded, the plate resistance will normally exceed 50,000 ohms. In the case of screen-grid tubes operated with the plate considerably more positive than the screen, the plate resistance is high, but less than in the corresponding radio-frequency pentode.

With low plate voltages the plate resistance of a pentode tube drops considerably and corresponds to the plate resistance of an ordinary triode tube. In a screen-grid tube with plate voltage less than the screen

voltage, the plate resistance will vary widely in value, as is apparent from the fact that the slope of the plate-current curve in Fig. 69 varies greatly. In particular it is to be noted that, when the secondary emission at the plate is considerable, the plate resistance may be negative.

Mutual Conductance (or Transconductance).—The mutual conductance (or, as it is sometimes called, the transconductance) of screen-grid and pentode tubes is defined in the same way as in triodes and is given by the equation

$$\text{Mutual conductance} = G_m = \frac{\partial I_p}{\partial E_g} = \frac{\mu}{R_p} \tag{69}$$

where

I_p = plate current
E_g = control-grid voltage
μ = amplification factor given by Eq. (67)
R_p = plate resistance as given by Eq. (68).

The mutual conductance represents the rate of change of plate current with control-grid voltage, or, what is the same thing, the ratio μ/R_p. The mutual conductance is the most important single constant of screen-grid and pentode tubes when operated in the usual manner with sufficient plate voltage to make the plate current substantially independent of plate voltage. Under such conditions the plate current is very nearly equal to the total space current, and is also proportional to the total space current. The mutual conductance will then depend primarily upon the magnitude of plate current, but not upon the combination of control-grid, screen, and plate potentials required to produce the current. The mutual conductance of screen-grid and pentode tubes under such conditions is analogous to the mutual conductance of triodes and, in fact, has about the same numerical value as the mutual conductance of a corresponding triode at the same plate current.

At plate voltages so low that the plate current is not independent of plate voltage, the mutual conductance depends very greatly upon the plate potential as well as upon the total space current, and will vary over wide limits with changes in plate voltage. In particular, with screen-grid tubes operated at potentials that give a negative plate resistance, the mutual conductance will be negative.

Miscellaneous Constants.—Screen-grid and pentode tubes possess numerous other constants which may under special circumstances be useful in expressing properties of the tube. Thus each positive electrode has its own dynamic resistance defined in the same way as the plate resistance except that the expression is in terms of the voltage and current of the electrode involved. Likewise, each positive electrode possesses a mutual conductance or transconductance with respect to every other electrode; in the general case this is defined by the relation

$$\text{Mutual conductance} = \frac{\partial I_1}{\partial E_2} \tag{70}$$

where I_1 is the current to the electrode for which one desires a mutual conductance and E_2 is the potential of the electrode with respect to which one desires the mutual conductance. Finally, there are numerous amplification factors, each of which is defined in terms of the relative effectiveness of some particular pair of electrodes upon some current in the tube. Thus in the pentode tube one could define amplification factors giving the relative effectiveness of control-grid and suppressor-grid potentials upon the plate current, upon the screen-grid current, and upon the total space current $I_p + I_{sg}$.

33a. The Beam Tube.—The beam tube is a special type of screen-grid tube in which the effect of a suppressor grid is obtained by means of the

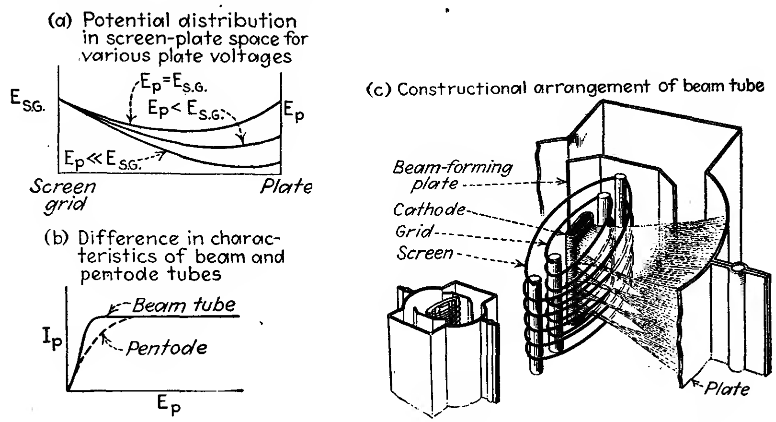


FIG. 73.—Characteristics and constructional features of beam power tube.

space charge in the space between plate and screen. This is accomplished by using a large plate-screen distance and by concentrating the electrons traveling toward the plate into a well-defined beam. These modifications increase the space-charge effects existing in the plate-screen space to the point where there is a pronounced potential minimum in the interelectrode space even when the plate potential is low. This potential minimum produced by the space charge is illustrated in Fig. 73*a* and performs the same function as a suppressor grid in that it causes secondary electrons emitted from the plate to encounter an opposing field which returns them to the plate. In spite of this potential minimum, the plate still collects all the electrons that pass the screen provided the minimum potential in the interelectrode space exceeds zero. The characteristic that results is shown in Fig. 73*b* and is the same as a pentode characteristic except that the transition between the region where the plate current is substantially

independent of plate voltage and where the plate current is determined primarily by plate voltage is more abrupt and takes place at a lower plate potential. The gradual transition of the pentode is caused by the fact that the spaces between the wires of the suppressor grid tend to produce a variable-mu action not present in the beam tube.

The construction of a practical beam tube is shown in Fig. 73*c*. The beam-forming plates are internally connected to the cathode and serve to concentrate the electrons in a beam as indicated. This increases the space-charge effect sufficiently so that, when coupled with the large screen-plate distance, the space charge will produce the required potential minimum. The beam-forming plates also serve to keep the electrons away from the ends of the control-grid structure where pronounced dissymmetry exists. The control-grid and screen-grid wires are carefully aligned so that the screen-grid wires are in the shadow cast by the control grid. This reduces to an unusually low value the fraction of the total space current which is intercepted by the screen.

The beam tube is equivalent to a pentode as far as fundamental characteristics are concerned, but its abrupt transition characteristic at low plate voltages makes it superior to the pentode as a power amplifier.

34. Variable-mu Tubes.[1]—Variable-mu tubes (also called remote cut-off tubes and supercontrol tubes) are screen-grid and pentode tubes in which the design has been modified in such a way as to cause the total space current of the tube to tail off at very negative control-grid potentials rather than to have a well-defined cut-off point. A typical characteristic of such a tube is shown in Fig. 74*a*, together with the characteristic of an ordinary screen-grid tube. A variable-mu characteristic is obtained by using a non-uniform control-grid structure, so that the amplification factor μ_{sg} is different for different parts of the tube. Such an arrangement makes the grid potential required for cut-off different for different parts of the tube, and those parts having the lowest value of μ_{sg} will hence require an extremely negative grid bias to cut off all plate current. The usual method of obtaining the variable μ_{sg} is illustrated in Fig. 74*b*, and consists in varying the pitch of the control-grid structure. The coefficients of variable-mu tubes are defined in the same way as for other pentode and screen-grid tubes, and, except for the variable character of μ_{sg} and the fact that at very low plate currents the mutual conductance changes less rapidly with grid bias, are similar in all respects to the constants of sharp cut-off screen-grid and pentode tubes.

Variable-mu tubes are used where it is desired to control the amplification by varying the control-grid potential of the tube. The characteristic

[1] See Stuart Ballantine and H. A. Snow, Reduction of Distortion and Cross-talk in Radio Receivers by Means of Variable-mu Tetrodes, *Proc. I.R.E.*, vol. 18, p. 2102, December, 1930.

curves of ordinary screen-grid and pentode tubes possess considerable curvature where the plate current (and hence mutual conductance) is low, whereas the characteristic curves of a variable-mu tube are only slightly curved under these same conditions, as is apparent from Fig. 74*a*. The low curvature of the variable-mu characteristic at low plate currents (and hence at low mutual conductance) minimizes cross-talk interference and distortion that would otherwise be produced.

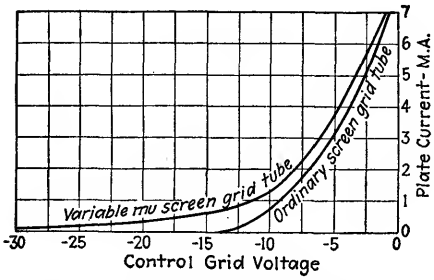


FIG. 74*a*.—Characteristic curve of a typical variable-mu tube compared with the characteristic curve of a corresponding tube of ordinary construction.

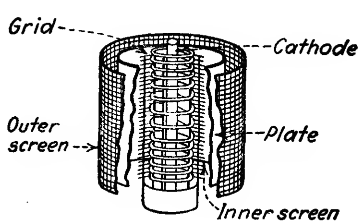


FIG. 74*b*.—Cut-away illustration of a variable-mu screen-grid tube, showing the variable pitch of the control-grid structure that gives the variable-mu characteristic.

35. Effect of Positive Control Grid.—Throughout the discussion given above for different types of tubes it has been assumed that the control grid is operated at a negative potential. If the control grid becomes positive, the total space current is still determined by the strength of the electrostatic field at the surface of the cathode, just as with the grid negative, but part of this current is diverted away from the other positive electrodes to the control grid. Hence Eqs. (59), (63), and (65) become

For triodes:

$$I_p + I_g = K\left(E_g + \frac{E_p}{\mu}\right)^{3/2} \tag{71}$$

For screen-grid and pentode tubes:

$$I_p + I_{sg} + I_g = K\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)^{3/2} \tag{72}$$

where I_g is the control-grid current and the remaining notation is the same as before.

The division of this total space current between the control grid and the remaining electrodes depends upon the type of tube and the electrode potentials. In the case of triodes the control-grid current increases as the control grid becomes more positive. The grid current is relatively small, however, and the rate of increase is slow until the grid voltage equals or

exceeds the plate potential, at which point the grid current suddenly begins to increase very rapidly. This is because the grid then attracts secondary electrons from the plate. The grid current also increases with

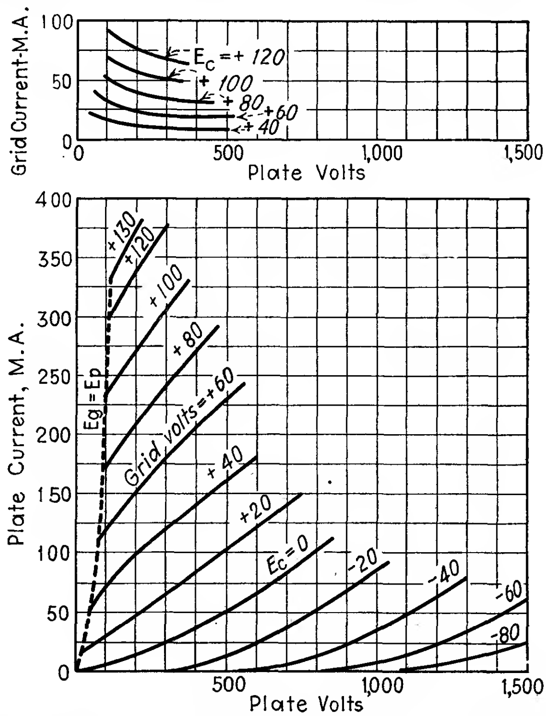


FIG. 75.—Characteristics of a small-power triode (type 800) in the positive control-grid region. It will be noted that the control-grid current tends to increase rapidly as the plate becomes less positive than the grid.

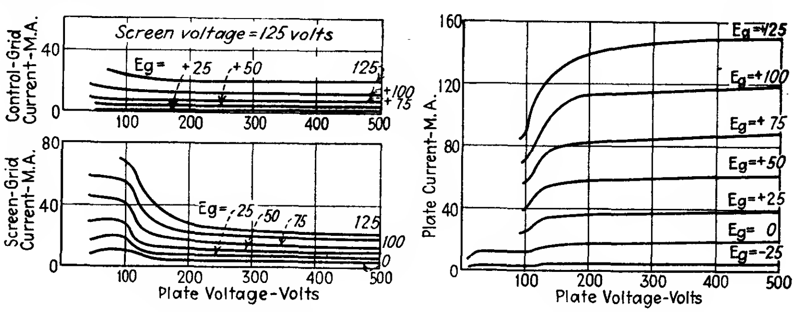


FIG. 76.—Characteristics of a small-power screen-grid tube (type 865) in the positive control-grid region.

reduction in plate voltage, particularly when the plate voltage is low enough to permit the formation in front of the plate of a virtual cathode that turns some electrons back toward the grid. The characteristics of a

triode tube in the positive grid region are of especial importance in connection with Class C amplifiers and power oscillators. A typical set of tube characteristics in this region is shown in Fig. 75. The control-grid current of screen-grid and pentode tubes increases with control-grid voltage, but tends to increase less rapidly at low plate voltages than in the case of triodes because the presence of the screen reduces the opportunity for the control grid to attract the secondary electrons produced at the plate. Characteristics of a screen-grid tube in the positive grid region are shown in Fig. 76.

36. Miscellaneous Tubes and Tube Applications.—The following paragraphs describe miscellaneous tube types and also a number of special circuit arrangements for standard triode, screen-grid, and pentode tubes.

Duplex Tubes.—A duplex tube is essentially two separate tubes in the same glass envelope. These tubes may be entirely independent, with separate leads for all electrodes, or they may be interrelated by having a common cathode. Duplex tubes are commonly used in radio receivers, and have the advantage of economy of space and a somewhat lower cost than two separate tubes to accomplish the same purpose. It is always possible, however, to use two tubes to accomplish the result of a single duplex tube.

Dual-grid or Class B Tubes.—Dual-grid or Class B tubes are characterized by the fact that they can be arranged to function as a triode tube which has either a very high or a moderately low amplification factor, and also by the fact that with the arrangement giving a high amplification factor the current that is drawn by the control grid when this grid is less positive than the plate is much smaller than is the case with an ordinary high-amplification-factor tube.

Dual-grid tubes consist of a cathode, two concentric grids, and a plate. For operation with low amplification factor, the outer grid is connected to the plate, while the inner grid serves as the control electrode. To obtain a high amplification factor the two grids are connected together and used as the control grid. This arrangement shields the cathode very effectively from the effect of a plate voltage, so that the amplification factor is extremely high. At the same time this high amplification factor is obtained with a relatively coarse structure for each grid, so that the total projected area of the grid wires is much smaller than the projected area of the wires of a single grid tube with sufficiently close mesh to give the same high amplification factor. As a consequence, the grid current of the dual-grid high-amplification-factor tube is unusually small when the grid is positive.

Dual-grid tubes are used in Class B audio amplifiers, and are discussed further in Sec. 61.

Space-charge-grid Tubes.—In space-charge-grid tubes there is an auxiliary grid called the space-charge grid located between the cathode and the control grid and operated at a low positive potential. The effect of the space-charge grid is to increase the number of electrons drawn out of the space charge near the cathode. Some of these electrons are immediately attracted to the space-charge grid, but many of them pass through its meshes into the space in front of the control grid, where they are slowed down by the retarding field and form a second space charge as shown in Fig. 77. This represents a virtual cathode which serves as the actual cathode for the remainder of the tube, which may be a triode, screen-grid tube, etc. The characteristic curves that result are similar to those for conventional tubes, as is apparent from Fig. 78.

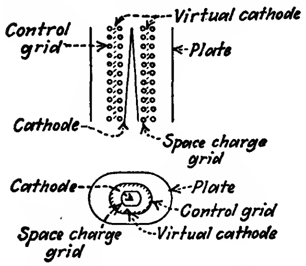


FIG. 77.—Details of space-charge grid tube. The inner grid is the space-charge grid and is operated at a moderate positive potential.

The advantages of the space-charge-grid arrangement arise from the fact that the virtual cathode has a large area and is located very close to the control grid. This gives a very high mutual conductance in proportion to the plate potential. The disadvantages of the arrangement are that the characteristic curves tend to have excessive curvature when considered over an appreciable range of voltages and that the space-charge grid draws a very heavy current, usually more than half of the total space current.

Any tube with more than one grid can ordinarily be arranged to function as a space-charge-grid tube. Thus the curves of Fig. 78 were obtained by using a conventional screen-grid tube with the ordinary control grid functioning as the space-charge grid and with the ordinary screen grid serving as the control grid. The result is then a space-charge-grid triode tube. A pentode tube can be likewise rearranged to serve as a space-charge-grid screen-grid tube, as in Fig. 79.

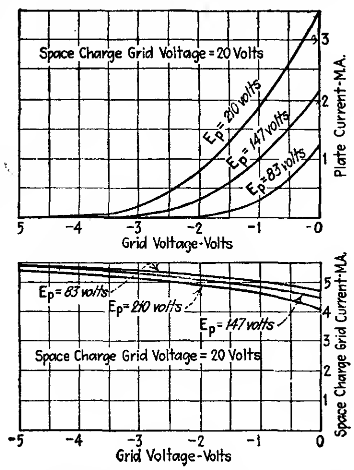


FIG. 78.—Characteristic curves of space-charge grid tube. The plate current varies with control-grid and plate voltages in much the same way as in a triode, while the space-charge grid current is much larger than the plate current and decreases as the plate current increases.

Special Connections for Conventional Tubes.—It is possible to operate conventional tubes to give special characteristics either by employing special connections or by using the proper combination of electrode

voltages. Several such examples have already been considered. Thus the dual-grid tube can be arranged as a triode having either a moderate or a high amplification factor, depending upon the way in which the grids are connected. Likewise, it was seen that any tube with two or more concentric grids could be made to operate as a space-charge-grid tube. The number of such arrangements is very great, particularly when the

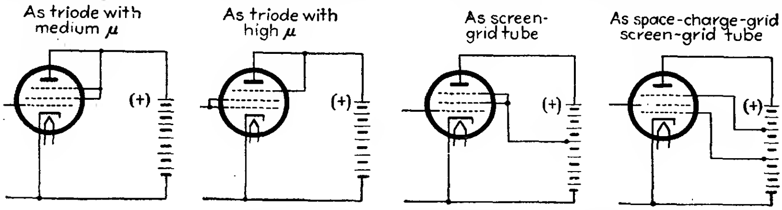


FIG. 79.—Pentode tube arranged in various ways.

number of electrodes at one's disposal becomes large. Thus an ordinary pentode tube can be connected as a triode having either a very high or a moderate amplification factor, as a screen-grid tube, or as a space-charge-grid screen-grid tube, as shown in Fig. 79.

Another way in which an ordinary tube can be rearranged is to interchange the functions of the grid and plate by making the grid the anode electrode and by using the plate as the negative control electrode, as shown in Fig. 80.[1] The operation of such an *inverted tube* rests on the fundamental fact that the space current flowing to the anode, which in this case is the positive grid, depends almost solely upon the electrostatic field in the vicinity of the cathode and is substantially independent of how this field is produced. Since both plate and grid potentials affect the intensity of this electrostatic field, it is possible to use a negative plate as a control electrode to serve the same purpose as the negative grid in the usual triode. The principal differences in the result are that the amplification factor is low, being approximately $1/\mu$, where μ is the amplification factor of the tube operated in the normal manner, and that the dynamic anode resistance is much lower than in the corresponding tube operated in the normal manner because changes in grid, *i.e.*, anode, voltage produce large changes in the electrostatic field near the cathode and hence large changes in anode, *i.e.*, grid, current. The inverted vacuum tube is a useful laboratory tool when it is necessary to control a current by a very high voltage without at the same time consuming any energy from the high potential source.

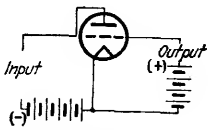


FIG. 80.—Circuit of inverted vacuum tube. The tube is an ordinary triode in which the plate is the control electrode and is operated at a negative potential, while the grid acts as the positive anode.

[1] See F. E. Terman, The Inverted Vacuum Tube, a Voltage-reducing Power Amplifier, *Proc. I.R.E.*, vol. 16, p. 447, April, 1928.

Coplanar Grid Tubes; Wunderlich Tube.—The coplanar grid tube can be thought of as an ordinary tube to which there has been added a second grid wound between the meshes of the usual grid. Such a double grid tube obviously introduces a number of possibilities not present in simpler tubes and has been found useful in a variety of circumstances.

If one of the coplanar grids is made positive, it is possible to produce a strong electrostatic field at the surface of the cathode even if the other grid is operated as a negative control grid, so that the result is a much larger space current than could be obtained in a corresponding triode tube. At the same time, if the plate is more positive than the positive coplanar grid, the latter will intercept only a small fraction of the total space current. A tube operated in this way has characteristics that are very suitable for power amplification.[1]

Another use of the coplanar grid tube is as a power grid-leak detector tube. This application is discussed in Sec. 85, and it is the purpose for which the Wunderlich coplanar grid tube was first developed, although this tube has since found a wide variety of other uses, mostly of a laboratory character.

Mixer Tubes for Superheterodyne Receivers.—Several types of multigrid tubes have been developed for use as the first detector of a superheterodyne radio receiver. The most important examples of these are the pentagrid converter tubes and the hexode mixer, both of which have five concentric grids. Inasmuch as the operation of these tubes is intimately related to the operation of the detector, they are considered in Sec. 88 in connection with detectors.

Tubes for High Frequencies.—At very high frequencies, such as 100 mc and more, ordinary tubes either become inoperative or function in a very unsatisfactory manner. This is partly due to the increasingly important role played by the interelectrode tube capacities and the inductance of the lead-in wires and partly due to the fact that at these high frequencies the length of time required for an electron to travel from the cathode to the plate is not negligible in comparison with the time represented by a cycle. The finite transit time causes changes in plate current to lag behind changes in grid potential. This gives rise to a number of very detrimental effects, such as cathode bombardment by electrons trapped in the interelectrode space at the instant the grid potential becomes greater than cut-off, and high grid loss. This grid loss occurs even when the grid is maintained negative, because of the fact that the variation in grid potential during the time an electron is in transit causes the grid and the electron stream to interchange energy. This effect is discussed further in Sec. 53.

[1] See H. A. Pidgeon and J. O. McNally, A Study of the Output Power Obtained from Vacuum Tubes of Different Types, *Proc. I.R.E.*, vol. 18, p. 266, February, 1930.

The natural solution for these limitations at high frequencies is to reduce the size of the tube as far as possible. It can be readily demonstrated that, if all dimensions are varied in the same proportion, the mutual conductance, amplification factor, and plate resistance will be unaffected but the interelectrode capacity, the lead inductance, and the transit time for the same electrode voltages will vary directly with the linear dimension. Miniature triode and pentode tubes (commonly called *acorn* tubes) based upon this principle are commercially available. These have substantially the same static characteristics as standard receiving tubes, but the detrimental effects of interelectrode capacity, lead inductance, and transit time are reduced by a factor of two to five over standard tubes.[1]

The disadvantage of small tubes is that the heat-dissipating ability of the anodes and the available thermionic emission both decrease with size, so that small tubes are inherently unsuited for handling appreciable power. Various expedients have been suggested for increasing the allowable dissipation at small anodes, and a number have been tried and found to show considerable improvement over standard tubes.[2] There is still room for improvement, however, and, although considerable progress has already been made, the ultimate solution of the ultra-high-frequency high-power tube does not appear to have been reached.

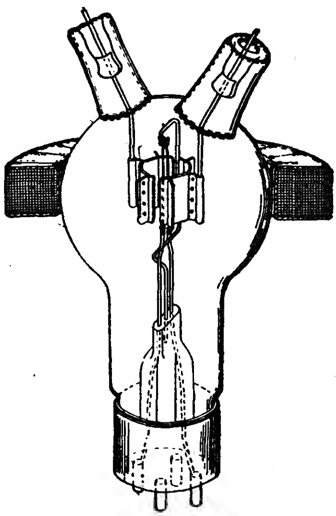


FIG. 81.—Split-anode magnetron. This consists of a filament surrounded by two semi-cylindrical anodes, and placed in a magnetic field so arranged that the flux lines are parallel to the filament.

Magnetrons.—A magnetron is a vacuum tube in which the flow of electrons from the cathode to the plate is affected by a magnetic field. A number of types of magnetrons have been suggested but the only one used to any extent in radio work is the split-anode magnetron illustrated in Fig. 81.[3] This consists of a filamentary cathode and two semi-

[1] See B. J. Thompson and G. M. Rose, Jr., Vacuum Tubes of Small Dimensions for Use at Extremely High Frequencies, *Proc. I.R.E.*, vol. 21, p. 1707, December, 1933; Bernard Salzberg and D. G. Burnside, Recent Developments in Miniature Tubes, *Proc. I.R.E.*, vol. 23, p. 1142, October, 1935.

[2] See C. E. Fay and A. L. Samuel, Vacuum Tubes for Generating Frequencies above One Hundred Megacycles, *Proc. I.R.E.*, vol. 23, p. 199, March, 1935.

[3] For information regarding other types of magnetrons, see Albert W. Hull, The Magnetron, *Jour. A.I.E.E.*, vol. 40, p. 715, September, 1921; The Axially Controlled Magnetron, *Trans. A.I.E.E.*, vol. 42, p. 915, 1923; Frank R. Elder, The Magnetron Amplifier and Power Oscillator, *Proc. I.R.E.*, vol. 13, p. 159, April, 1925.

cylindrical plates, with an axial magnetic field as illustrated. When the plates are positive, the electrons that are attracted to them follow curved paths as in Fig. 82. If the plates are very positive, the curvature of these paths is small, but at lower voltages the curvature becomes greater, until at a critical potential, determined by the strength of the magnetic field, the electrons follow a curved path (cardiod) back to the cathode, as shown in Fig. 82*c*, and never get to the plate in spite of the positive plate potential.

One of the chief reasons for the usefulness of the magnetron arises from the fact that, if the magnetic-field strength is slightly less than the cut-off value, and if a voltage is applied between the anodes in such a way as to make one anode less positive as the other is made more positive, the anodes will be found to offer a *negative* resistance to this superimposed

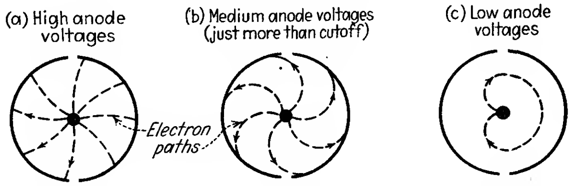


FIG. 82.—Electron paths in split-anode magnetron under various conditions.

potential. This negative resistance arises from the fact that the electrons which are attracted from the cathode by the most positive plate follow such a curved path as actually to reach the least positive plate, and vice versa, so that the added current which the electrode receives as a result of its added potential has the opposite sign from the potential increment.

Vacuum Tubes as Negative-resistance Devices.[1]—There are a number of ways in which a vacuum tube can be connected to give a negative resistance. The best known arrangement of this type is the dynatron, in which the negative resistance is obtained as a result of secondary emission at the plate as explained in Sec. 32. The dynatron is essentially a screen-grid tube operated with the plate less positive than the screen and having appreciable secondary emission at the plate. The plate circuit of the tube then offers a negative resistance having a magnitude that can be varied by the control-grid potential, as is apparent from Fig. 69. The lowest negative resistance obtainable from commercial screen-grid tubes used as dynatrons is 10,000 to 20,000 ohms.

Another convenient method of obtaining a negative resistance is to connect a pentode tube as shown in Fig. 83.[2] Here the screen-grid

[1] For further information about the dynatron and its uses see Albert W. Hull, The Dynatron, a Vacuum Tube Possessing a Negative Resistance, *Proc. I.R.E.*, vol. 6, p. 5, February, 1918.

[2] See E. W. Herold, Negative Resistance and Devices for Obtaining It, *Proc. I.R.E.*, vol. 23, p. 1201, October, 1935.

and plate electrodes are both positive, and the suppressor is sufficiently negative to produce a virtual cathode between the suppressor and screen. A high resistance is placed between the suppressor and its bias potential, and the suppressor and screen electrodes are tied together with a by-pass condenser. With this arrangement the screen-grid circuit offers a negative resistance to alternating currents. This is because the screen and suppressor electrodes are at the same potential with respect to alternating currents, and, if the potentials of both electrodes vary together, the current to the screen will decrease with increasing potential, and vice versa. The reduction in screen current comes about because the screen current consists largely of electrons that have returned from the virtual cathode produced by the negative suppressor. As the suppressor-grid potential increases, more electrons are drawn from the virtual cathode to the plate, leaving fewer to return to the screen and hence reducing the screen current. The magnitude of the negative resistance obtained in this manner can be controlled by varying the control-grid potential. With ordinary tubes values as low as 3000 or 4000 ohms are obtainable.

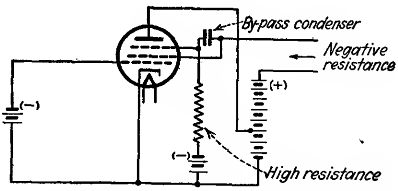


FIG. 83.—Retarding field method of producing a negative resistance.

A number of other negative resistance arrangements using vacuum tubes have been proposed but none of these except the magnetron, which is discussed above, has had much practical use.[1]

Negative resistances obtained from tubes can be used as ordinary circuit elements, and make it possible to achieve circuit behaviors not realizable with positive circuit elements. Thus it is possible to devise amplifiers, oscillators, etc., based upon circuits containing negative resistance elements,[2] although none of these arrangements except the dynatron oscillator (see Sec. 69) have been used to any extent.

37. The Mathematical Representation of Characteristic Curves of Tubes.—In carrying out the analysis of circuits involving vacuum tubes, it is often desirable to be able to express the characteristic curves of the tubes by means of a mathematical expression. The principal methods that have been employed to do this are the power-law method and the power-series method.

Power-law Method of Expressing Tube Characteristics.—This method of representing tube characteristics has already been made use of in Eqs.

[1] A very thorough discussion of various means of using tubes to obtain a negative resistance is given by E. W. Herold, *loc. cit.*

[2] See A. W. Hull, *loc. cit.;* E. W. Herold, *loc. cit.;* L. C. Verman, Negative Circuit Constants, *Proc. I.R.E.*, vol. 19, p. 676, April, 1931.

(57), (59), (63), (65), (71), and (72), which for the sake of convenience will be rewritten below:

For diodes:

$$I_p = KE_p^{3/2} \tag{73}$$

For triodes:

$$I_p + I_g = K\left(E_g + \frac{E_p}{\mu}\right)^{3/2} \tag{74}$$

For screen-grid and pentode tubes:

$$I_p + I_{sg} + I_g = K\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)^{3/2} \tag{75}$$

Also in pentodes and screen-grid tubes, when the plate potential is sufficiently high to make the plate current substantially independent of plate voltage, the plate current is almost exactly proportional to the total space current, so that for these conditions one has

$$I_p = K\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)^{3/2} \tag{76}$$

The notation in these equations is the same as that used when they were first developed.

It is assumed in Eqs. (73) to (76) that the velocity with which the electrons are emitted from the cathode and also the contact potentials in the grid and plate circuits are negligibly small, and that there is no voltage drop in the filament. Equations (74), (75), and (76) also assume that the amplification factors μ and μ_{sg} are geometrical constants entirely independent of electrode voltages, which is equivalent to assuming that the control-grid structure has perfect symmetry. It is also assumed that there is a full space charge about the cathode and that an electron which has once been drawn out of this space charge will not return to it. Hence the equations do not necessarily hold when a virtual cathode is formed somewhere within the tube.

The effect of the voltage drop in the filament of filament-type tubes and also the effects of velocity of emission and contact potentials can be taken into account in Eqs. (73) to (76) by methods discussed when these relations were first developed in Secs. 27, 28, 31, and 32. These corrections are quite small except when the field near the cathode is very weak, *i.e.*, when the space current is small. When the corrections are made, Eq. (73) gives a very accurate representation of the complete characteristic of a diode tube throughout the region where a full space charge exists. The remaining equations are still only approximate representations, however, because the lack of perfect symmetry in actual tubes causes the amplification factors μ and μ_{sg} to depend somewhat upon the electrode voltages.

Because of this the characteristics of triodes, screen-grid tubes, and pentodes can usually be more accurately represented by equations of the following types:

For triodes:

$$I_p + I_g = K\left(E_g + \frac{E_p}{\mu}\right)^{\alpha} \tag{77}$$

For screen-grid and pentode tubes:

$$I_p + I_{sg} + I_g = K\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)^{\alpha} \tag{78}$$

When the plate voltage is great enough so that the plate receives most of the space current, then Eq. (78) becomes:

$$I_p = K\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)^{\alpha} \tag{79}$$

In these equations the amplification factor μ or μ_{sg}, as the case may be, is evaluated for a point near the center of the region in which the most accurate representation is desired, and the exponent α is determined experimentally by plotting a curve of current against the parenthesis on the right-hand side of the equations, using logarithmic paper as shown in Fig. 84. The value of α at any current is the slope of the resulting curve at that point, and, although the slope is not perfectly constant, a particular value of α will give a good approximation over an appreciable range.

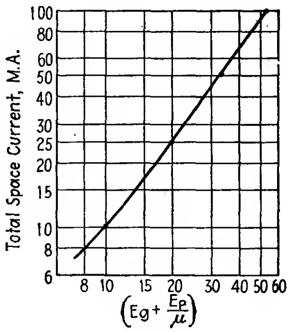


FIG. 84.—Plot of total space current as a function of $(E_g + E_p/\mu)$ for a triode, showing how to evaluate α in the equation $I_p = K(E_g + E_p/\mu)^{\alpha}$.

The power-law method of representing tube characteristics will be applied in Secs. 61 and 63 to the analysis of Class C amplifiers and harmonic generators.

Power-series Method of Representing Characteristic Curves of Tubes.—In the power-series method, the tube characteristics are expressed in terms of a Taylor's series, or power series.[1] The details of this method can be understood by applying it to the case of a triode with equipotential cathode and assuming that over the limited range that is to be represented the amplification factor μ can be considered constant. The method can then be extended to the general case in which variations in the amplification factor are taken into account.

[1] This method of analysis was first proposed by Carson; see John R. Carson, A Theoretical Study of the Three-element Vacuum Tube, *Proc. I.R.E.*, vol. 7, p. 187, 1919.

For the simplified triode case the plate current is some function of the quantity $\left(E_g + \frac{E_p}{\mu}\right)$. That is:

$$I_p = f\left(E_g + \frac{E_p}{\mu}\right) \tag{80}$$

This equation can be corrected for contact potential and velocity of emission in the usual manner, if desired. In obtaining results based upon Eq. (80), one is normally interested, not in the total plate and grid voltages and total plate current, but rather in variations of the plate current which result from variations of the electrode voltages about some operating point corresponding to a plate voltage E_{p0}, grid voltage E_{g0}, and a plate current I_0. Hence one can write:

$$\left.\begin{aligned}\text{Actual plate current} &= I_p = I_0 + i_p\\ \text{Actual plate voltage} &= E_p = E_{p0} + e_p\\ \text{Actual grid voltage} &= E_g = E_{g0} + e_g\end{aligned}\right\} \tag{81}$$

The lower-case letters represent the variations from the operating point, such as might result when a small signal voltage is applied to the grid of the vacuum tube and produces changes in the plate current and plate voltage. A substitution of Eqs. (81) into Eq. (80), and then expanding the latter into a Taylor series gives

$$\begin{aligned}i_p = f\left(e_g + \frac{e_p}{\mu}\right) &= \frac{1}{1!}\left(\frac{\partial I_p}{\partial E_g}\right)_0\left(e_g + \frac{e_p}{\mu}\right) + \frac{1}{2!}\left(\frac{\partial^2 I_p}{\partial E_g{}^2}\right)_0\left(e_g + \frac{e_p}{\mu}\right)^2\\ &+ \frac{1}{3!}\left(\frac{\partial^3 I_p}{\partial E_g{}^3}\right)_0\left(e_g + \frac{e_p}{\mu}\right)^3 + \cdots + \frac{1}{k!}\left(\frac{\partial^k I_p}{\partial E_g{}^k}\right)_0\left(e_g + \frac{e_p}{\mu}\right)^k\end{aligned} \tag{82}$$

The subscript 0 denotes that the derivatives are to be evaluated at the point $E_g = E_{g0}$, $E_p = E_{p0}$, and $I_p = I_{p0}$ of the characteristic.

The physical significance of Eq. (82) can be made clearer by noting that

$$\left.\begin{aligned}\frac{\partial I_p}{\partial E_g} &= G_m = \frac{\mu}{R_p}\\ \frac{\partial^2 I_p}{\partial E_g{}^2} &= \frac{\partial G_m}{\partial E_g} = \frac{-\mu}{R_p{}^2}\frac{\partial R_p}{\partial E_g}\\ \frac{\partial^3 I_p}{\partial E_g{}^3} &= \frac{\partial^2 G_m}{\partial E_g{}^2} = \frac{2\mu}{R_p{}^3}\left(\frac{\partial R_p}{\partial E_g}\right)^2 - \frac{\mu}{R_p{}^2}\frac{\partial^2 R_p}{\partial E_g{}^2}\end{aligned}\right. \tag{83}$$

where R_p is the plate resistance, G_m is the mutual conductance, and all derivatives are evaluated at the point E_{p0}, E_{g0}, I_{p0}. Substituting the relations of Eq. (83) into Eq. (82) gives

$$i_p = G_m\left(e_g + \frac{e_p}{\mu}\right) + \frac{1}{2!}\left(\frac{\partial G_m}{\partial E_g}\right)\left(e_g + \frac{e_p}{\mu}\right)^2 + \frac{1}{3!}\left(\frac{\partial^2 G_m}{\mu E_g^2}\right)\left(e_g + \frac{e_p}{\mu}\right)^3 + \cdots \tag{84}$$

$$= \frac{\mu}{R_p}\left(e_g + \frac{e_p}{\mu}\right) - \frac{\mu}{2R_p^2}\left(\frac{\partial R_p}{\partial E_g}\right)\left(e_g + \frac{e_p}{\mu}\right)^2 + \cdots \tag{85}$$

Equations (82) and (84) express the characteristics of the vacuum tube about the operating point E_{p0}, E_{g0}, and I_{p0} at which the tube characteristics are evaluated, and they are exact provided a sufficient number of terms are included in the series and provided the plate current I_p is greater than zero and is less than the saturation value. In actual practice the series converges very rapidly so that two or at most three terms are sufficient to explain all important aspects of tube behavior.

A more general form of Eqs. (82) and (84) can be obtained by expressing the plate current in the form $I_p = f(E_g, E_p)$. Substituting Eqs. (83) into this expression gives a result that can be expanded as a Taylor series of two variables.[1] The method can also be extended to include screen-grid and pentode tubes.[2]

The power-series method of representing tube characteristics finds use in the analysis of the factors causing amplitude distortion in amplifiers, in the analysis of detection, and in certain other phenomena such as automatic synchronization. Applications of the method to these specific problems are taken up in later chapters.

38. Effect of Gas upon Tube Characteristics.—Very small traces of gas in vacuum tubes affect the characteristics adversely in a number of ways as a result of the positive ions produced in the tube by collisions between the gas molecules and the electrons flowing to the plate. The positive ions travel in the opposite direction from electrons and normally end their existence by falling into the cathode or the negative control grid. Those which bombard the cathode tend to destroy the emission of thoriated-tungsten and oxide-coated cathodes, as has already been discussed, and this effect is sufficiently serious to limit the usefulness of these types of emitters. The positive ions collected by the negative control grid also result in grid current that causes grid-circuit power loss and limits the resistance that may be inserted in series with the grid, as discussed below. Another serious effect of the positive ions is the irregularities that they produce in the space charge at the cathode. This

[1] This analysis, taking into account the variation of the amplification factor and also extended to include the effect of control-grid current, is due to F. B. Llewellyn, Operation of Thermionic Vacuum-tube Circuits, *Bell System Tech. Jour.*, vol. 5, p. 433, July, 1926.

[2] See J. G. Brainerd, Mathematical Theory of Four-electrode Tubes, *Proc. I.R.E.*, vol. 17, p. 1006, June, 1929.

causes a shot effect that makes the tube "noisy"; this will be discussed in Sec. 49.

When traces of gas are present in a tube, the resistance that may safely be placed in series with the negative control grid of the tube is limited. This is because the voltage drop that the grid current produces across such a resistance has a polarity that makes the grid less negative. This increases the total space current, thereby increasing the number of positive ions and causing additional grid current and a still greater reduction in the negative grid potential. If the resistance in the grid circuit is high enough, this process can become cumulative, resulting in the control-grid potential suddenly becoming positive and causing the destruction of the tube as a result of excessive plate current. As a consequence there is a maximum resistance that it is permissible to place in series with the grid electrode, with the allowable value depending upon the tube characteristics and the conditions under which the tube is operated.

When the gas pressure is increased somewhat, the voltage and current relations become seriously affected. The positive ions in drifting toward the cathode neutralize a portion of the negative space charge around the cathode and make it possible for more electrons to be drawn to the plate than would otherwise be possible. The velocity of the positive ions is so low because of their large mass that the life of the average ion is much greater than that of an electron. Hence the rate of production of positive ions need not be very great to increase the plate current appreciably. The characteristic curves of tubes containing small amounts of gas also tend to be irregular and to have sudden bends or kinks.

When the amount of gas in a tube is quite large, the ionization is sufficiently intense to produce a luminous glow, and the tube is then said to be "soft." The amount of gas required to produce a soft tube is quite small, since a pressure of one-millionth of an atmosphere will commonly give a visible glow and make the tube inoperative as a high-vacuum device.

The amount of gas in a tube can be estimated from the amount of grid current present when the grid is negative. This current is proportional to the number of positive ions, which, with appreciable anode voltages, are proportional to the total space current and the gas pressure. This is the basis of the ionization gauge used to measure very low gas pressures.

Hot-cathode Gas Diodes.—When the gas pressure in a tube is of the order of 1 to 30 μ, as is the case when the tube contains mercury vapor in equilibrium with liquid mercury at room temperatures, the presence of the gas profoundly affects the characteristics. In the case of a diode, the plate current starts to increase with plate voltage in exactly the same

way as in the high-vacuum tube of Fig. 51, but at some critical potential, which is just above the ionizing potential of the gas, there is a sudden break and the current increases to the full cathode emission with little or no increase in the plate voltage, as shown in Fig. 85. This is caused by the appearance of positive ions, which at these gas pressures will be produced in sufficient quantities at plate voltages just above the ionizing potential to neutralize completely the space charge of the negative electrons around the cathode. The result is that, as soon as positive ionization sets in, the full emission current can be drawn to the plate with just enough plate potential to keep the ionization process functioning. This sort of characteristic finds practical application in the hot-cathode mercury-vapor rectifier tubes and is discussed in Sec. 94.

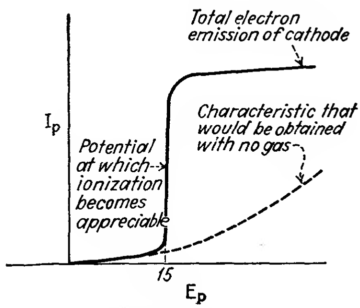


FIG. 85.—Characteristic of hot-cathode mercury-vapor diode, showing how the space-charge limitation to the space current is removed as soon as ionization begins.

Gas Triodes (Thyratrons).—When gas having a pressure of the order of 1 to 30 μ is introduced into a triode (or any tube having a control grid), the control action exerted by the grid is changed in a very remarkable way. If one starts with a grid potential considerably more negative than the cutoff value, and then gradually reduces this negative bias, it is found that, at the point where the plate current would just start to flow if the tube contained no gas, the plate current suddenly jumps from zero to a very high value, which readily reaches the full emission of the cathode with anode voltages as low as 15 to 20 volts. After the flow of plate current has once been started, the control grid has no further effect, and the grid can be made much more negative than cutoff without altering the plate current appreciably. To stop the plate current one must reduce the plate voltage below the ionizing potential of the gas in the tube.

The above characteristic is caused by the fact that, as soon as the plate current starts to flow, positive ions are produced as a result of ionization by collision. Some of these are attracted toward the negative grid, surrounding it with a sheath of positive ions that neutralize the electrostatic effect of the grid and so destroy the normal control action of the grid. At the same time other positive ions are attracted toward the cathode and neutralize the space charge. Hence, once ionization has started, there is no space charge to limit the current flow, and the control action of the grid has been lost.

The result is a relay or trigger device that has numerous important practical uses, particularly in control work. It takes practically no

energy at the negative grid to initiate the discharge, and at the same time the resulting energy turned on may be many kilowatts. Tubes of this character are known as gas triodes, or thyratrons, or grid-glow tubes. When such tubes are to be used in low-frequency circuits, as in association with 60-cycle power sources, the gas employed is usually mercury vapor in equilibrium with liquid mercury. In other cases where the ionization and deionization must be extremely rapid, helium, argon, or neon is employed because the positive ions of these gases are much lighter and hence more mobile than mercury ions.[1]

39. Constructional Features of Small Tubes.—The following paragraphs describe very briefly the principal constructional features of the types of tubes commonly used in radio receivers. The details of larger tubes are discussed in Chap. VII in connection with power amplifiers.

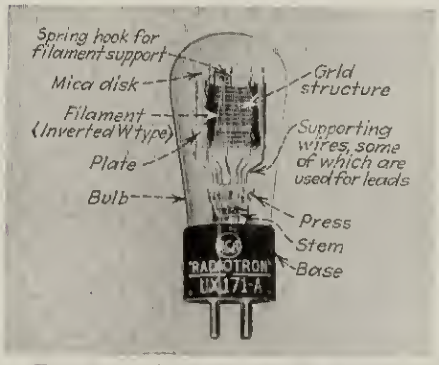


FIG. 86*a*.—Photograph showing constructional details of typical three-electrode vacuum tube.

Glass-envelope Tubes.—The outstanding features of glass-envelope tubes are shown by Figs. 86*a* and 86*b*. The electrodes are supported from wires embedded in the press, with certain of the wires passing through the glass and acting as leads. There is also usually some method, commonly a mica sheet, for holding the tops of the electrodes in proper position with respect to each other. The technique of assembling a tube is similar to that employed in the manufacture of incandescent lights. The electrodes are formed to the proper shape and are spot-welded to the supporting wires, which are then held in position while the press is made at the end of the stem. Next the bulb is sealed to the stem, after which the tube is evacuated and based.

The plates of small tubes are usually formed of nickel or iron.[2] The most common construction uses sheet material in which the rigidity is often increased by such expedients as crimping the flat surface and turning over the edges to form a flange. The surface of the plate is sometimes blackened by carbonization to facilitate the radiation of the heat produced by the impact of electrons.

The grid structure is preferably of molybdenum, but less expensive materials such as nichrome, iron, iron-nickel alloys, and manganese-nickel

[1] For further information on gas triodes and their uses, see Keith Henney, "Electron Tubes in Industry," pp. 144–272, McGraw-Hill Book Company, Inc.

[2] An extensive discussion of materials entering into the manufacture of receiving tubes is given by E. R. Wagner, Raw Materials in Vacuum-tube Manufacture, *Electronics*, vol. 7, p. 104, April, 1934.

alloys are commonly used. In the usual method of construction the grid is attached to its supporting wires by being wound in grooves cut in the support wires at suitable intervals, after which a swaging process is used to hold the grid wires in place mechanically.

The filament in filament-type tubes is strung in the form of an inverted V or an inverted W upon hooks which keep it under tension. In the case of heater-type tubes the heaters are usually wound in a double spiral which is slipped into the hollow cathode sleeve after being first sprayed with an

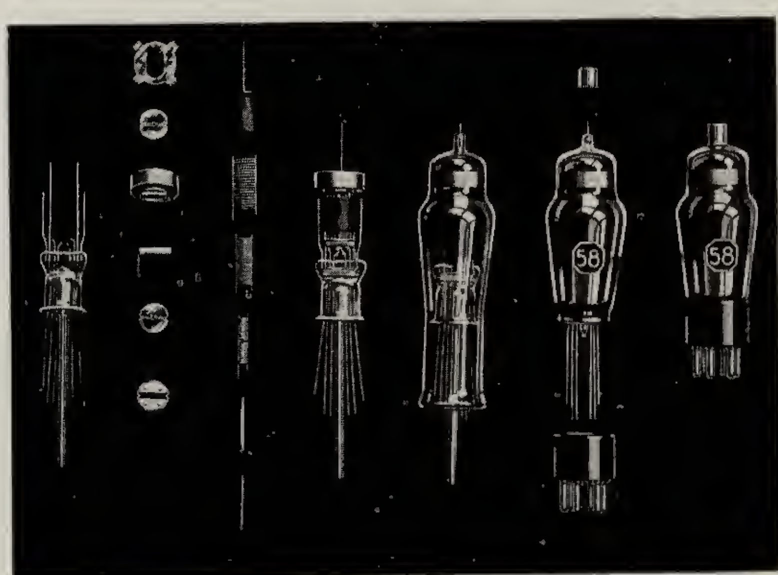


FIG. 86*b*.—Photograph showing constructional features of a heater-type variable-mu pentode.

insulating compound that prevents short circuits between the heater turns or between the heater and cathode sleeve.

Metal-envelope Tubes.—The metal-envelope tube differs from the glass tube primarily in that the outer envelope consists of a metal shell rather than a glass bulb. This change alters the manufacturing technique in many important respects, although the internal assembly of cathode, grids, and plate is essentially the same irrespective of the type of envelope employed.

The principal features of a metal-envelope tube are shown in Fig. 87. Instead of being mounted upon the press, the various electrodes are supported from a metal "header," as shown. The lead-in wires are brought out through glass beads that are sealed to fernico or kovar eyelets, which in turn are welded to the header. Fernico is a nickel-iron alloy which has the same coefficient of expansion as the glass bead and so makes possible a

vacuum-tight joint. After the electrode structure is completely assembled upon the header, the metal shell is welded to the header as indicated in Fig. 87, and the tube is finally evacuated, sealed off, and based. The evacuation is accomplished through a metal exhaust tube welded to the header, and the sealing-off process consists in pinching this tube and welding it closed.

As compared with glass tubes, the metal tube has certain advantages. It is probable that better and at the same time less expensive tubes are possible by using the metal construction. In particular the elimination of

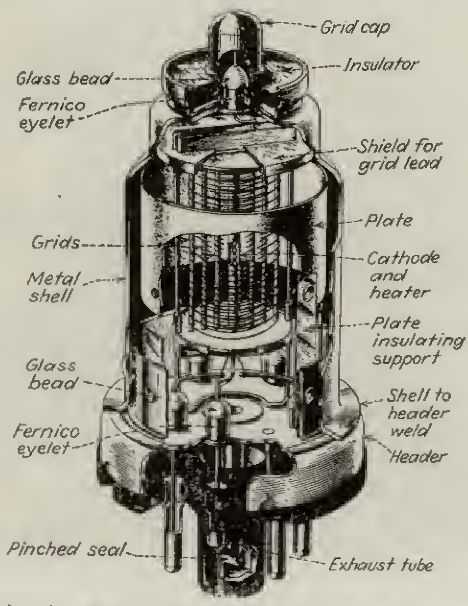


FIG. 87.—Cut-away drawing showing constructional features of a metal-envelope pentode tube.

the glass-blowing technique will probably make it possible to maintain improved tolerances in clearances.

Evacuation of Tubes.[1]—The air is removed from the tube by the use of a motor-driven oil-immersed pump supplemented by a molecular pump. If the required degree of vacuum is to be maintained for any length of time, it is necessary to remove the gas adsorbed and absorbed by the metal and glass parts as well as that contained in the space inclosed by the envelope. In the case of glass-envelope tubes, the removal of these occluded gases is carried out by baking the entire tube at a temperature just below the softening point of glass and by heating the electrodes to a high tempera-

[1] A discussion of some of the practical considerations involved in the evacuation of small tubes is given by E. R. Wagner, Processes in Vacuum-tube Manufacture, *Electronics*, vol. 7, p. 213, July, 1934.

ture, usually by means of a radio-frequency furnace. As a final step, a small quantity of some substance called a "getter" is volatilized (or flashed) inside the tube to remove the residual gas either by chemical or by mechanical action. Magnesium is widely used for this purpose, but other materials such as barium, phosphorus, etc., can also be employed, as well as various mixtures. Some getters also act as "keepers" in that they not only remove what gas is present in the tube at the time of flashing but also combine with any gas that may subsequently be liberated within the tube.

With metal-envelope tubes, the occluded gases are removed by heating the metal shell with a gas flame. Gas can be removed from the plate,

TABLE V.—CHARACTERISTICS OF TYPICAL VACUUM TUBES USED FOR VOLTAGE AMPLIFICATION

Type	Equivalent tubes*	Heater or filament			Normal electrode rating					Properties		
		Type	Voltage	Current	Plate volts	Grid volts	Screen volts	Screen ma	Plate ma	G_m	R_p	μ
						Triode Tubes						
56	76	Heater	2.5	1.00	250	−13.5	.....	...	5.0	1,450	9,500	13.8
30	....	Filament	2.0	0.06	180	−13.5	.....	...	3.1	900	10,300	9.3
2A6§	75§, 6Q7§†	Heater	2.5	0.8	250	− 2.0	.....	...	0.8	1,100	91,000	100
6C5†	....	Heater	6.3	0.3	250	− 8.0	.....	...	8.0	2,000	10,000	20
6F5†	....	Heater	6.3	0.3	250	− 2.0	.....	...	0.9	1,500	66,000	100
55§	85§, 27	Heater	2.5	1.0	250	−20.0	.....	...	8.0	1,100	7,500	8.3
						Screen-grid Tubes						
24A	....	Heater	2.5	1.75	250	−3	90	1.7	4.0	1,050	0.6 meg.	630
32	....	Filament	2.0	0.06	180	−3	67.5	0.4	1.7	650	1.2 meg.	780
35‡	....	Heater	2.5	1.75	250	−3	90	2.5	6.5	1,050	0.4 meg.	420
						Radio-frequency Pentode Tubes						
6C6	57, 6J7†, 77	Heater	6.3	0.3	250	−3	100	0.5	2.0	1,225	1.5 meg.	1,500
6D6‡	78‡, 6K7‡†, 58‡	Heater	6.3	0.3	250	−3	100	2.0	8.2	1,600	0.8 meg.	1,280
34‖‡	....	Filament	2.0	0.06	180	−3	67.5	1.0	2.8	620	1.0 meg.	620
39/44‖	....	Heater	6.3	0.3	250	−3	90	1.4	5.8	1,050	1.0 meg.	1,050
6B7§	2B7§	Heater	6.3	0.3	250	−3	125	2.3	9.0	1,125	0.65 meg.	730

* These tubes have similar (but not necessarily identical) Normal Electrode Ratings and Properties, and are suitable for performing the same functions.

† Metal envelope.

‡ Variable mu.

§ Tube with diode sections.

‖ Suppressor grid internally connected to cathode.

grids, etc., by heating these electrodes, using electron bombardment to supplement the general heating of the outer envelope.

40. Table of Tube Characteristics.—The essential characteristics of different types of vacuum tubes in common use are given in Tables V, VIII, IX, XI, and XII. The first table covers tubes used for voltage amplification and other similar purposes. These tubes may be classified as triodes, radio-frequency pentodes, and screen-grid tubes, and they may be either of the sharp-cut-off or variable-mu types. Small-power tubes such as are used in radio receivers are listed in Table VIII (Chap. VII), while large-power tubes, such as are used in radio transmitters, are covered in Table IX (Chap. VII). Rectifier tubes are given in Tables XI and XII (Chap. XI).

Problems

1. In a two-electrode tube in which the cathode and plate electrodes can be considered as plane surfaces 1 cm apart, the anode voltage is +200 volts.

a. With what velocity do the emitted electrons strike the anode?

b. How long does it take an electron to travel from cathode to anode if there is a uniform voltage gradient (*i.e.*, if the space charge is negligible)?

2. In Prob. 1 there is an anode current of 100 ma. How many electrons arrive at the anode each second and what is the energy dissipation at the anode?

3. In the tube of Prob. 1 it will be found that, when there is a strong magnetic field between the electrodes oriented so that the flux lines are parallel with the plane electrodes, the anode current becomes zero, but, if the flux lines are perpendicular to the plane electrodes, there is little or no effect on the plate current. Explain.

4. In an oxide-coated cathode operating at 1150°K. and having $A = 0.005$, calculate the thermionic emission per square centimeter when the constant b has values 12,000 and 11,000.

5. A certain water-cooled tube has a tungsten filament 19.5 in. long and 0.025 in. in diameter.

a. What electron emission can be expected at 2450°K. and at 2600°K.?

b. What is the ratio of filament-heating powers required at the two temperatures?

6. Calculate and plot a curve giving relative electron emission as a function of temperature for a typical oxide-coated emitter.

7. Discuss the economic factors involved in the selection of the cathode operating temperature in a vacuum tube.

8. In tubes employing thoriated-tungsten emitters it is found that the immediate effect of an accidental overheating is to cause the electron emission to drop to a low value, but that after further operation at normal or slightly above normal temperature the emission is gradually restored to normal. Explain, and also explain why oxide-coated and pure tungsten emitters do not act in this manner.

9. Explain why heater-type cathodes always use oxide-coated emitters.

10. In a two-electrode tube it is found that at a plate voltage of +100 the plate current is 90 ma, while saturation begins to be pronounced at a plate voltage of +200 volts.

a. From this information calculate the plate current as a function of plate voltage where space-charge limitation exists, plotting the resulting curve and also sketching in the saturation region.

b. Calculate the plate voltage at which saturation would begin to occur if the electron emission were reduced to one-third of its initial value by lowering the cathode temperature, and sketch a dotted curve alongside of (*a*) to show the effect of the change.

11. Derive Eqs. (58*a*) and (58*b*).

12. Using the plate-current–plate-voltage characteristic curve of a two-electrode tube obtained from a tube manual (or assigned), check the extent to which Eq. (57) holds by plotting I_p versus E_p on log paper.

13. Derive equations analogous to Eqs. (58*a*) and (58*b*), but taking into account the effect of voltage drop in the filament of a filament-type triode tube.

14. Using the characteristic curves of a triode tube (either the tube of Fig. 54 or a tube from a tube manual), check the extent to which Eq. (59) holds by plotting I_p versus $\left(E_g + \frac{E_p}{\mu}\right)$ on log paper.

15. Explain how one could plot an entire family of grid-voltage–plate-current curves such as shown in Fig. 54, knowing the amplification factor of the tube and having available one curve of the family.

16. Evaluate μ, G_m, and R_p for the tube of Figs. 54 and 55 in the region of $E_p = 135$, $I_p = 17$ ma.

17. Two identical triode tubes are connected in parallel. How do μ, G_m, and R_p of the combination compare with the values for the individual tubes.

18. Assuming Eq. (59) holds, derive an expression giving the way in which the mutual conductance of a triode can be expected to vary with total space current.

19. In a particular triode tube it is found by experiment that $\mu = 10$ and $K = 0.0005$. From this information:

a. Calculate and plot a family of $E_g - I_p$ curves similar to those of Fig. 54.

b. Calculate and plot R_p and G_m as a function of total space current.

20. If the tube of Prob. 19 is placed in parallel with a second tube for which $\mu = 20$ and $K = 0.0005$.

a. Calculate and plot a curve giving the resultant I_p of the combination as a function of E_g for a plate voltage of 250 volts, and compare the shape with the shape of a corresponding curve of Prob. 19.

b. Calculate and plot R_p, G_m, and μ of the combination as a function of total space current, when the total space current is varied by means of the grid bias and when the anode voltage is 250 volts.

c. Repeat (*b*) for a plate potential of 150 volts.

d. Discuss the results of (*b*) and (*c*) as compared with the results of Prob. 19*b*, and from this draw conclusions as to the effect of dissymmetries in a tube which cause different parts to have different amplification factors.

21. Explain why the suppressor grid in an ordinary pentode tube has a relatively coarse mesh compared with the screen and control grids (see Fig. 86*b*).

22. What would be the effect on the curves of Figs. 63 and 65 if the filament temperature were reduced to the point where saturation occurred at about 5 ma total space current?

23. If you were given a tube that might be either a pentode with the suppressor internally connected to the cathode or a screen-grid tube, but you did not know which and could not see through the glass bulb well enough to make sure, what electrical tests could be used to determine which kind of tube it was?

24. Under what conditions might the voltage and current relations of a screen electrode of a screen-grid tube show a negative resistance characteristic?

25. Discuss the effect that secondary emission at the screen grid has on the coefficients of screen-grid and pentode tubes.

26. Evaluate the coefficients G_m, R_p, μ, and μ_{sg} of the pentode tube of Figs. 63 and 64 in the vicinity of $E_g = -1.5$, $E_{sg} = 100$, $E_p = 150$.

27. Under some conditions it is found that the current drawn by a positive control grid is negative. Explain how this can happen in a high-vacuum tube and state the conditions favorable to the production of a negative control-grid current.

28. There are no variable-mu triodes manufactured. Explain how to obtain a tube having the characteristics of a variable-mu triode by a proper connection of the electrodes of some standard tube that is manufactured.

29. Describe how a screen-grid tube could be connected to function as: (*a*) high-mu triode; (*b*) low-mu triode; (*c*) space-charge-grid tube; (*d*) inverted tube.

30. Can a coplanar grid tube be connected to function: (*a*) as a triode; (*b*) as a space-charge-grid tube; (*c*) as a screen-grid tube?

31. Evaluate the negative resistance in the tube of Fig. 69 when $E_g = 0$, $E_{sg} = 83$, and $E_p = 40$.

32. *a*. To what extent can Eqs. (77), (78), and (79) be used to express the characteristics of variable-mu tubes?

b. To what extent can the power series of Eq. (82) be used to express the characteristics of variable-mu tubes?

33. *a*. Evaluate the first three coefficients of Eq. (82) in terms of I_p, K, and μ, assuming that the plate current accurately follows Eq. (59).

b. If $K = 0.001$ and $\mu = 13$, evaluate the first three coefficients for $I_p = 0.005$ amp.

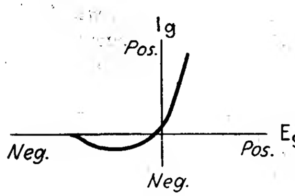


FIG. A.

34. When a tube contains traces of gas, the grid current varies with grid potential as shown in Fig. *A* when the voltage of the plate and other electrodes is constant. Explain the reason for this type of curve.

35. In evacuating tubes, the gas pressure is ordinarily determined by sealing a triode tube to the exhaust manifold so that the gas pressure inside the tube is the same as that in the manifold. The tube elements are connected to form an inverted vacuum tube, and the gas pressure is evaluated in terms of the current to the negative plate. (*a*) State how this determines the gas pressure. (*b*) Explain how the plate current could be expected to vary with total space current and with gas pressure. (*c*) Would the nature of the gas within the tube have any effect on the calibration?

CHAPTER V

VACUUM-TUBE AMPLIFIERS

41. Vacuum-tube Amplifiers.—The amplifying property of a vacuum tube results from the fact that the control grid draws no electrons when maintained at a negative potential, while at the same time variations in the negative grid potential cause corresponding variations in the space current. In this way it is possible for a voltage representing practically no energy to control a relatively large space current and to develop a

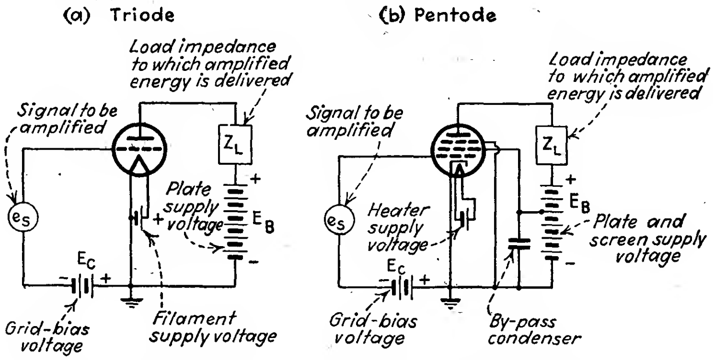


FIG. 88.—Basic circuit of vacuum-tube amplifiers employing triode and pentode tubes.

corresponding quantity of electrical energy in an anode circuit. The vacuum tube is capable of amplifying voltages of all frequencies up to the highest used in radio communication, and by employing a number of tubes in cascade almost any desired amount of amplification can be obtained. The vacuum-tube amplifier is therefore one of the most important instruments in electrical engineering, having made possible the long-distance telephone, talking pictures, radio sets with loud-speakers, television, etc.

The basic circuits of two typical amplifiers are shown in Fig. 88. The voltage E_c is for the purpose of maintaining the control grid negative with respect to the cathode so that no electrons will be drawn to the grid; it is known as the grid bias or C-voltage. The signal to be amplified is represented in the figure by e_s and is applied to the grid in series with the grid-bias voltage. Under ordinary conditions the amplitude of signal is so related to the grid-bias voltage that the instantaneous control-grid potential never becomes positive with respect to the cathode. Variations in

the plate current that results from the action of the signal voltage on the control grid must flow through the impedance Z_L, which is in series with the plate supply voltage and is known as the load impedance. The energy that is supplied to the load impedance by the variations in the plate current represents the useful part of the amplified energy. The signal voltage applied to the control grid supplies substantially no energy to the tube when the grid is negative because a negative grid draws no electrons, but at the same time it causes plate-current variations that deliver considerable quantities of energy to the load impedance. The result is that

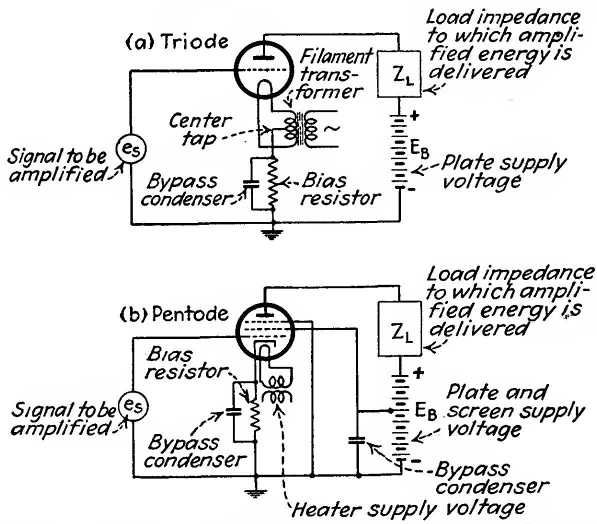


FIG. 89.—Vacuum-tube amplifiers similar to those of Fig. 88 except that the control grid is made negative with respect to the cathode by a self-bias resistance instead of a battery.

the output energy of the tube, *i.e.*, the energy delivered to the load impedance, can be made many times that required to produce the signal applied to the grid.

In some circumstances the control-grid potential of an amplifier is allowed to go moderately positive at the positive crests of the signal voltage. This causes the grid to attract some electrons, and means that the signal to be amplified must supply some energy to the tube. However, if the grid is not allowed to go too far positive, the number of electrons attracted by the control grid is relatively small, and the energy that the signal is required to supply is still much less than the energy developed in the load impedance. Hence there is still amplification of energy, although to a lesser extent than when the control grid is always negative.

In practical amplifiers it is usually desirable to obtain the bias voltage from the source of plate potential rather than from a separate battery. This can be accomplished, as shown in Fig. 89, by connecting a resistance

between the cathode and ground and by-passing this resistance with a condenser large enough to be an effective short circuit to alternating currents of the lowest frequency to be amplified. Such an arrangement places the cathode at a positive potential with respect to the control grid, and so gives the equivalent of a negative bias for the grid. The amount of bias obtained is the voltage drop of the *total space current* flowing through the bias resistance, and is controlled by the amount of resistance. Further discussion of bias arrangements is given in Sec. 93.

Methods of Classifying Amplifiers.—Amplifiers are classified in ways descriptive of their character and properties. The first classification is according to the frequency to be amplified, and leads to the four broad divisions known as audio-frequency, radio-frequency, video-frequency, and direct-current amplifiers. Audio-frequency amplifiers are intended for amplifying currents of audible frequencies, *i.e.*, from about 15 cycles per second to approximately 10,000 cycles. Frequencies higher than 10,000 to 15,000 cycles per second are considered as radio frequencies, while video frequencies are those contained in television signals, and range from 5 to 10 cycles up to over 1,000,000 cycles.

Amplifiers are also classified as to whether they handle a wide or narrow band of frequencies. A band of frequencies is considered wide or narrow in proportion to the ratio of the width of the band to the frequency at its center. Thus the group of frequencies lying between 100 and 5000 cycles is said to represent a wide band, while the frequency band from 1,000,100 to 1,005,000 cycles, which extends over the same frequency range, is narrow. When substantially equal amplification is to be obtained over a band that is wide according to this definition, the amplifier is said to be untuned, while, when the amplifier is associated with sharply resonant circuits so arranged that only a narrow band of frequencies is amplified, it is spoken of as a tuned amplifier.

Amplifiers can also be divided into voltage and power amplifiers according to whether the object is to produce as much voltage or as much power as possible in the load impedance. The reason that it is necessary to distinguish between these objectives can be understood by considering the problem of obtaining a large quantity of undistorted power output from a signal voltage that is so small as to require more amplification than can be obtained from a single tube. Under such circumstances it is customary to use a number of amplifiers in cascade, each one of which amplifies the output of the preceding tube and delivers its output to another tube for additional amplification. In arranging such an amplifier, the best results are obtained by making all the amplifying tubes except the last one operate as voltage amplifiers, while the last tube functions as a power amplifier. The power tube then has the maximum possible signal voltage applied to its grid and is therefore able to deliver the greatest

amount of power. There is no object in making the intermediate tubes function as power amplifiers, since the purpose of these tubes is to increase the signal voltage delivered to the last tube, the power output of which represents the output of the amplifier.

Amplifiers, particularly power amplifiers, are also designated as Class A, Class AB, Class B, Class C, or linear amplifiers according to the operating conditions. The term Class A is applied to an amplifier adjusted so that the plate current flows continuously throughout the cycle of the applied signal voltage. In the Class B and linear amplifiers the tube is biased approximately to cut-off so that the plate current in an individual tube flows in pulses lasting approximately a half cycle, while the Class C amplifier is adjusted so the plate current flows in pulses that last less than a half cycle. The Class AB is intermediate between the Class A and Class B adjustments. The linear amplifier and Class B amplifier differ only in that the former employes a tuned load circuit.

The type of amplifier is fixed primarily by the constants of the associated electrical circuits and by the grid and plate voltages employed. It is possible to make any particular tube function as any kind of amplifier, although the tube characteristics best suited for each type of amplifier are somewhat different.

42. Distortion in Amplifiers.—An ideal amplifier produces an output that exactly duplicates the input in all respects except magnitude. An actual amplifier can fall short of this ideal by failing to amplify the different frequency components of the input voltage equally well, by giving an output that is not exactly proportional to the amplitude of the input, or by making the relative phases of the different frequency components in the output differ from the relative phase relations existing in the input. These effects are commonly referred to as *frequency*, *amplitude* (or non-linear), and *phase distortion*, respectively.

Frequency distortion (*i.e.*, unequal amplification of different frequencies contained in the signal) is particularly important in audio- and video-frequency amplifiers and is more difficult to eliminate the wider the band of frequencies that must be amplified with substantial equality. Frequency distortion tends to be greater as the amplification per stage is increased, so if low distortion is desired it is necessary to make some sacrifice in the amplification. An example of frequency distortion is given in Fig. 90, in which the high-frequency component of the voltage to be amplified is discriminated against.

Amplitude distortion results from a non-linear relation between voltage and current in either the input (*i.e.*, grid) or the output (*i.e.*, plate) circuits of the amplifier. Non-linearity of the input circuit results when the grid is allowed to go positive during a portion of the signal-voltage cycle. This is because the grid-cathode resistance of the tube is

relatively low when the grid is positive, and extremely high when negative, with the result that the voltage actually applied to the grid tends to be distorted if the grid is not maintained negative at all times. Amplitude distortion in the output circuit occurs as a result of the non-linear relation between grid voltage and plate current and, while not completely avoidable, can be kept small by proper attention to the conditions under which the amplifier operates. In particular, the amplitude distortion introduced by the plate circuit will be least at points where the plate current-plate voltage characteristic as given in Fig. 54 has the least curvature, and when the signal voltage applied to the grid is small.

Amplitude or non-linear distortion results in the production of frequencies in the amplifier output that are not present in the input voltage applied to the grid. The most important of these distortion frequencies are harmonics of the frequencies contained in the input, and sum and difference frequencies formed by combinations of the signal components. An example of amplitude distortion is shown in Fig. 90.

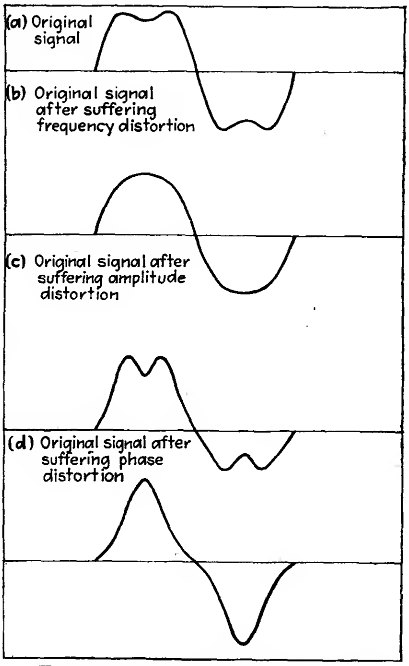


FIG. 90.—Series of waves showing the effects produced by frequency, amplitude, and phase distortion.

Phase distortion results when the relative phase relations of the different frequency components contained in the signal voltage are disturbed in such a way as to make the wave form of the amplifier output differ from the wave shape of the signal without changing the magnitudes of the frequency components that are involved. This effect is illustrated by the top and bottom waves in Fig. 90, which contain the same percentage of third harmonic in different phase relations with respect to the fundamental. The resulting wave forms are seen to be very different, although the amount of energy in the harmonic is the same in both cases. In order that the original wave shape may be maintained, it is necessary that the phase shift in the amplifier be exactly proportional to the frequency, and furthermore that the phase shift when extrapolated to zero frequency be zero or some integral multiple of π. The phase shift of an amplifier is related to the time required for the transmission of different frequencies through the amplifier, and when phase

distortion exists it means that different frequencies are transmitted with different speeds and hence do not arrive at the output at the same time. From a practical point of view, phase distortion is relatively unimportant in ordinary audio-frequency amplifiers because the phases can be altered over a wide range without producing an effect noticeable to the ear. It is only where the time of transmission has an order of magnitude comparable with the duration of the signal voltage applied that phase distortion becomes of importance. This situation is encountered in television circuits and in long telephone and telegraph lines.

43. Equivalent Circuit of the Vacuum-tube Amplifier.—The variations produced in the plate current of a vacuum tube by the application of a

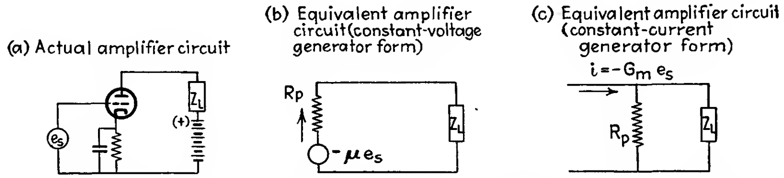


FIG. 91.—Equivalent circuits of the vacuum-tube amplifier. In the constant-voltage form the effect that is produced in the plate circuit by the signal e_s acting on the grid is taken into account by postulating that the plate circuit can be replaced by an equivalent generator of voltage $-\mu e_s$ acting from cathode toward the plate and having an internal resistance equal to the plate resistance R_p. In the constant-current form the tube is considered as generating a current $-G_m e_s$ acting from cathode toward the plate and flowing through the impedance formed by the plate resistance of the tube in parallel with the load impedance.

signal voltage to the control grid are exactly the same variations that would be produced in the plate current by a generator developing a voltage $-\mu e_s$ acting inside the tube from cathode toward the plate, in a circuit consisting of the tube plate resistance in series with the load impedance. The effect on the plate current of applying a signal voltage e_s to the grid is therefore exactly as though the plate-cathode circuit of the tube were a generator developing a voltage $-\mu e_s$ and having an internal resistance equal to the plate resistance of the tube. This leads to the equivalent circuit of the vacuum-tube amplifier, which is shown in Fig. 91(*b*) and which is the basis of most amplifier designs and calculations.[1]

[1] A rigorous proof of the equivalent amplifier circuit of the vacuum tube is given in Sec. 55. A somewhat less general demonstration is as follows: The plate current *flowing from cathode to plate* which a grid-voltage increment e_s produces will be called i_p. This plate current i_p causes a voltage drop $-Z_L i_p$ in the plate load impedance Z_L, so that the application of e_s to the grid reduces the voltage actually applied to the plate of the tube by an amount $\Delta E_p = -Z_L i_p$, and the plate-current increment i_p that flows is the result of the joint action of the voltage increment e_s applied to the grid and the reduction $-Z_L i_p$ in the plate voltage. Since the voltage e_s produces the same effect on the electrostatic field adjacent to the cathode as an added plate voltage of μe_s, the current i_p is the same current that would result from an increment in the plate voltage of $(\mu e_s + Z_L i_p)$, and when this increment is small the current

According to this equivalent circuit, it is apparent that the load current i_p produced by the application of a signal potential e_s to the control grid is

$$i_p = \frac{-\mu e_s}{R_p + Z_L} \tag{86a}$$

The signs are such that a positive value for i_p means a current flowing in opposition to the steady direct current present in the plate circuit when no signal is applied. The voltage developed across the load Z_L by the current i_p is

$$\text{Voltage across load} = i_p Z_L = \frac{-\mu e_s Z_L}{R_p + Z_L} \tag{86b}$$

It is sometimes convenient to rearrange Eqs. (86a) and (86b) by dividing both numerator and denominator of the right hand side by R_p. Doing this gives

$$i_p = \frac{-\mu}{R_p} \frac{e_s}{1 + \dfrac{Z_L}{R_p}} = -G_m e_s \frac{R_p}{R_p + Z_L} \tag{87a}$$

$$\text{Voltage across load} = -G_m e_s \frac{R_p Z_L}{R_p + Z_L} \tag{87b}$$

Here $G_m = \mu/R_p$ is the mutual conductance of the tube. This form of the tube equation shows that the effect of applying a signal voltage e_s to the control grid is the same as though the tube generated a current $-e_s G_m$ flowing from plate toward cathode through an impedance formed by the plate resistance in parallel with the load impedance. This leads to an alternative form of equivalent circuit, which is shown in Fig. 91c and which can be called the *constant-current generator* form of the equivalent circuit, as contrasted with the *constant-voltage generator* form shown in Fig. 91b.

The constant-current and constant-voltage forms of the equivalent circuit of the vacuum-tube amplifier lead to the same result as far as the load is concerned, and are, in fact, merely different ways of expressing the same relationship. The constant-voltage circuit of Fig. 91b is the most

---

increment flowing toward the plate is the negative of the plate-voltage increment divided by the dynamic plate resistance R_p, so that

$$i_p = -\frac{(\mu e_s + Z_L i_p)}{R_p}$$

or

$$i_p = \frac{-\mu e_s}{R_p + Z_L}$$

This last equation is simply a mathematical statement of the equivalent amplifier circuit of Fig. 91b.

convenient when triodes are involved, but with pentode and screen-grid tubes, where the plate resistance and amplification factors are very high, the constant-current arrangement of Fig. 91*c* is simpler to use.

The equivalent circuits of the amplifier give only those currents and voltage drops that are produced as a result of the application of a signal voltage upon the amplifier grid. The actual currents and potentials existing in the plate circuit are the sum of the currents and potentials developed in the equivalent circuit and those existing in the amplifier when no signal is applied. Since the steady values that are present when no signal is applied are of no particular interest so far as the amplifier performance is concerned, it is usually unnecessary to superimpose them upon the results calculated on the basis of the equivalent circuit. In particular, when the signal applied to the grid is an alternating voltage, as is usually the case, the equivalent circuit gives directly the a-c currents produced in the plate circuit by the signal voltage, which are the currents superimposed upon the direct-current quantities present when no signal is applied.

The equivalent circuit gives the exact performance of the vacuum-tube amplifier to the extent that the plate resistance R_p and the amplification factor μ, which are used in setting up the equivalent circuit, are constant over the range of variations produced in control-grid and plate voltages by the signal voltage. Hence, when the signal is small, the equivalent circuit is almost exactly correct because the changes produced by the signal are so small that the tube constants are substantially constant. As the signal voltage increases, the error involved in the equivalent circuit becomes larger. To obtain the *exact* behavior one must then modify the equivalent circuit to take into account the effects introduced by the variation in the circuit constants, as discussed in Sec. 55. However, for practical conditions the effects resulting from variations in the tube constants are second-order effects even with rather large signal voltages, so that the equivalent circuit is found useful for all signal amplitudes.

44. Audio-frequency Voltage Amplifiers. Resistance Coupling.—In voltage amplifiers the object is to obtain from the amplifier output as much voltage as possible to be applied to the grid of the succeeding amplifier tube. One way of doing this is to place in the amplifier plate circuit a high-resistance load called the coupling resistance.

The circuits of practical resistance-coupled amplifiers using triode and pentode tubes are shown in Fig. 92*a*, in which R_c is the resistance load across which the amplified voltage is developed. The grid-leak resistance and the coupling or blocking condenser shown in Fig. 92 are for the purpose of preventing the direct-current voltage applied to the plate of the amplifier tube from also being applied to the grid of the tube to which the amplified voltage is delivered. The coupling condenser should be

large enough to offer a low reactance to the frequencies to be amplified, while the grid leak should have a very high resistance in order that the shunting effect of the grid leak and coupling condenser upon the coupling resistance may be small.

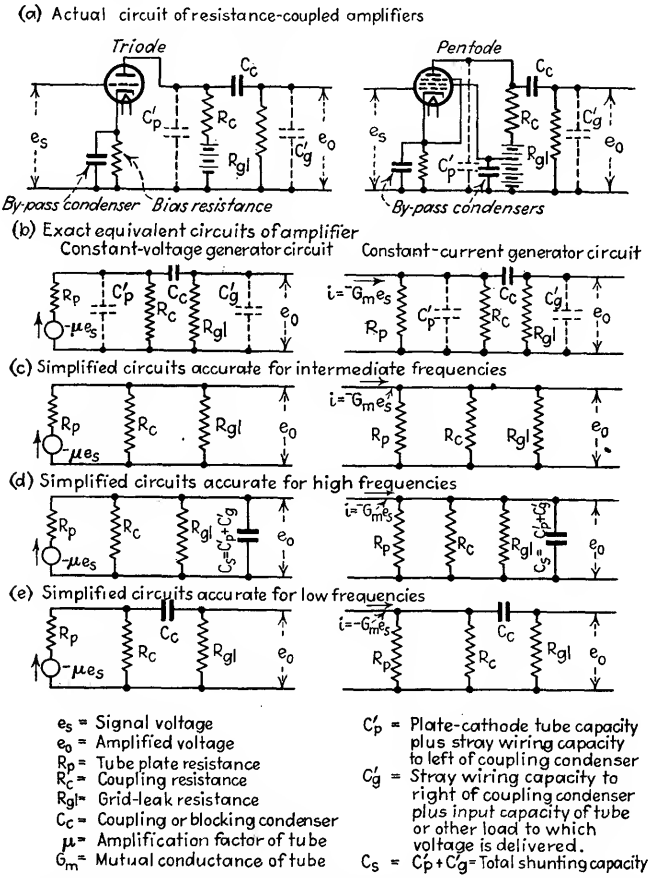


FIG. 92.—Circuits of resistance-coupled amplifiers, together with equivalent circuits and simplifications of the equivalent circuits useful in making amplifier calculations. The constant-voltage generator circuit is used with the triode tube, and the constant-current generator form of the equivalent circuit is used for the pentode tube.

The most important property of the resistance-coupled amplifier is the way in which the amplification varies with frequency. Such a characteristic is shown in Fig. 93 for a representative case and has as its distinguishing feature an amplification that is substantially constant over a wide range of frequencies, but which drops off at both very low and very high frequencies. The falling off at very low frequencies is a result of the fact that the high reactance which the coupling condenser C_c offers to low

frequencies consumes some of the low-frequency voltage that would otherwise be developed across the grid leak. The reduction in amplification at high frequencies is caused by the tube and stray capacities which shunt the coupling and grid-leak resistances, and which have low enough reactance at high frequencies to lower the effective load impedance, with a consequent reduction in the voltage developed at the output.

Analysis and Calculation of Amplification Characteristic.—In order to calculate the amplification and the way in which the amplification varies with frequency, the proper procedure is first to replace the tube by its equivalent electric circuit. This takes care of the tube characteristics and reduces the problem to the analysis of an ordinary electric circuit.

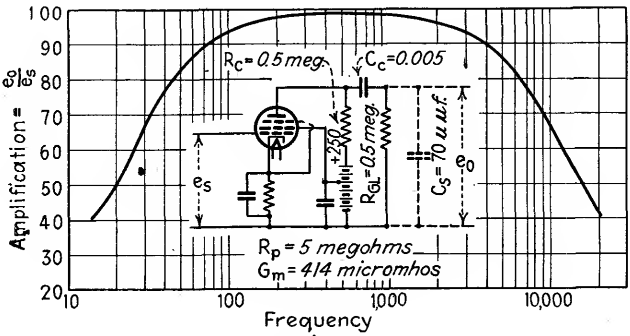


FIG. 93.—Variation of amplification with frequency in a typical resistance-coupled pentode amplifier.

Either form of equivalent circuit can be employed, with the constant-voltage arrangement most convenient for triode tubes and the constant-current arrangement more satisfactory for pentode and other tubes having very high plate resistances.

Equivalent circuits of a resistance-coupled amplifier showing all essential circuit elements are given in Fig. 92*b* for both methods of representing the tube characteristics. The resulting circuits are quite complicated, but can be simplified by considering only a limited range of frequencies at a time. Thus in the middle range of frequencies the reactance of the coupling condenser C_c in a properly designed amplifier will be so small as to be the practical equivalent of a short circuit as compared with the grid-leak resistance, whereas the reactance of the shunting capacities will still be so high as to be the practical equivalent of an open circuit.[1] Under such conditions the equivalent circuits take the form

[1] These assumptions fail only when the ratio of shunting to coupling capacities is unusually large, *i.e.*, when the coupling capacity is especially small and the shunting capacity is at the same time rather large. The error involved hence tends to become less the larger the frequency range which is to be amplified, and in practical cases it is negligible with amplifiers covering the audio-frequency range. Thus in the ampli-

shown in Fig. 92*c*. A very simple calculation based upon these circuits shows that the amplification in the middle range of frequencies is given by the equations:

For constant-voltage generator circuit (most suitable for triodes):

$$\left.\begin{array}{l}\text{Amplification in middle}\\ \text{range of frequencies}\end{array}\right\} = \frac{e_0}{e_s} = \mu\frac{R_L}{R_L + R_p} \tag{88a}$$

For constant-current generator (most suitable for pentodes).

$$\left.\begin{array}{l}\text{Amplification in middle}\\ \text{range of frequencies}\end{array}\right\} = \frac{e_0}{e_s} = G_mR_{eq} \tag{88b}$$

In these equations μ is the amplification factor of the tube as defined in Secs. 30 and 33, R_p is the plate resistance, R_L is the equivalent load resistance formed by the coupling resistance and the grid-leak resistance in parallel $\left(R_L = \dfrac{R_cR_{gl}}{R_c + R_{gl}}\right)$, while R_{eq} is the equivalent resistance formed by the plate resistance of the tube, the grid-leak resistance, and the coupling resistance, all in parallel.[1] That is

$$\begin{aligned} R_{eq} &= \frac{R_LR_p}{R_L + R_p} = \frac{R_cR_pR_{gl}}{R_cR_{gl} + R_cR_p + R_{gl}R_p} \\ &= \frac{R_c}{1 + \dfrac{R_c}{R_{gl}} + \dfrac{R_c}{R_p}} \end{aligned} \tag{89}$$

At high frequencies it is necessary to take into account the effect of the capacities shunting the coupling and grid-leak resistances, leading to the equivalent circuit of Fig. 92*d*, which is applicable at high frequen-

---

fier of Fig. 93 the neglect of C_c and C_s in the middle-frequency range introduces an error of about 0.5 per cent in the amplification. The ratio C_s/C_c in this case is $^{70}\!/_{5000} = 0.014$, which is larger, and hence more likely to give error, than the average design of resistance-coupled amplifier.

[1] It is to be noted that in the case of pentode tubes, where the plate resistance is considerably larger than the coupling and grid-leak resistances, only small error is introduced by assuming that

$$R_{eq} = R_L = \frac{R_cR_{gl}}{R_c + R_{gl}}$$

Further, if the grid-leak resistance is much larger than the coupling resistance, as is often the case, then, to a good approximation, $R_{eq} = R_L = R_c$, and the middle-range amplification becomes G_mR_c.

cies. A manipulation of the voltage and current relations involved gives the result[1]

$$\frac{\left.\begin{matrix}\text{Actual amplification}\\ \text{at high frequencies}\end{matrix}\right\}}{\left.\begin{matrix}\text{Amplification in}\\ \text{middle range}\end{matrix}\right\}} = \frac{1}{\sqrt{1 + (R_{eq}/X_s)^2}} \tag{90}$$

where

$X_s = 1/2\pi f C_s$ = reactance of total shunting capacity C_s

R_{eq} = resistance formed by plate, coupling, and grid-leak resistances, all in parallel as given by Eq. (89).

The extent to which the amplification falls off at high frequencies is therefore determined by the ratio which the reactance of the shunting capacity C_s bears to the equivalent resistance obtained by combining the plate resistance, coupling resistance, and grid-leak resistance in parallel. This loss of amplification at high frequencies can be estimated by the fact that at the frequency which makes the reactance of the shunting condenser C_s equal the equivalent resistance formed by R_p, R_c, and R_{gl} in parallel, the amplification drops to 70.7 per cent of its middle-frequency range value. The amount of falling off at other frequencies, conveniently related to the 70.7 per cent case, is given in Table VI.

At low frequencies the shunting capacity C_s has such a high reactance as to be equivalent to an open circuit, but the reactance of the coupling condenser C_c becomes sufficient to cause a falling off in the amplification.

[1] Equation (90) is derived by applying Thévenin's theorem to the network to the left of the shunting capacity C_s in Fig. 92*d*. According to Thévenin's theorem this network can be replaced by an equivalent generator in series with a resistance, as shown in Fig. 94. The voltage of the generator is the voltage appearing across the

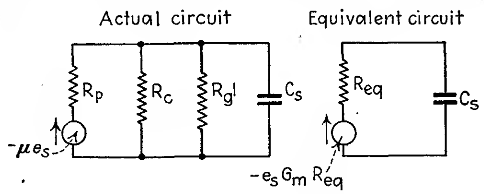


FIG. 94.—Simplification of the circuit of Fig. 92*d* by the use of Thévenin's theorem.

condenser terminals when the condenser is open-circuited, and so is the output voltage in the middle range of frequencies as given by Eqs. (88*a*) and (88*b*). The internal resistance of the generator is the resistance formed by plate, coupling, and grid-leak resistances, all in parallel, and so is given by Eq. (89). Referring to Fig. 94, it is apparent that the amplification at high frequencies is

$$\text{High-frequency amplification} = G_m R_{eq} \frac{X_s}{\sqrt{X_s^2 + R_{eq}^2}} = G_m R_{eq} \frac{1}{\sqrt{1 + (R_{eq}/X_s)^2}}$$

Dividing this by Eq. (88*b*) gives Eq. (90).

The equivalent circuit under these conditions hence takes the form shown in Fig. 92*e*. A manipulation of the relations existing in this circuit shows that

$$\frac{\left.\begin{matrix}\text{Amplification at}\\ \text{low frequencies}\end{matrix}\right\}}{\left.\begin{matrix}\text{Amplification in}\\ \text{middle range}\end{matrix}\right\}} = \frac{1}{\sqrt{1 + (X_c/R)^2}} \tag{91}$$

where

$X_c = 1/2\pi f C_c$ = reactance of coupling condenser C_c

$R = R_{gl} + \dfrac{R_c R_p}{R_c + R_p} = R_{gl} + \dfrac{R_c}{1 + \dfrac{R_c}{R_p}}$ = resistance formed by grid leak in series with the combination of plate and coupling resistances in parallel.

TABLE VI.—RELATIVE AMPLIFICATION OF RESISTANCE-COUPLED AMPLIFIERS

Low frequencies		High frequencies	
Frequency	Relative amplification	Frequency	Relative amplification
$5f_o$	0.980	$0.2f_o'$	0.980
$2f_o$	0.895	$0.5f_o'$	0.895
f_o	0.707	f_o'	0.707
$0.5f_o$	0.447	$2f_o'$	0.447
$0.2f_o$	0.196	$5f_o'$	0.196
$0.1f_o$	0.100	$10f_o'$	0.100

f_o = frequency at which coupling-condenser reactance equals the resistance formed by grid leak in series with parallel combination of plate and coupling resistances.

f_o' = frequency at which reactance of shunting capacity equals the resistance formed by plate, coupling, and grid-leak resistances, all in parallel.

The extent to which the amplification falls off at low frequencies is therefore determined by the ratio of the reactance of the coupling condenser to the equivalent resistance obtained by combining the grid leak in series with the parallel combination of coupling resistance and plate resistance. This loss of amplification at low frequencies can be estimated by the fact that, at the frequency that makes the reactance of the coupling condenser equal the equivalent resistance R, the amplification falls to 70.7 per cent of its value in the middle range of frequencies. The amount of falling off at other frequencies conveniently related to the 70.7 per cent case is given in Table VI.

Universal Amplification Curve of Resistance-coupled Amplifier.—The simplicity of Eqs. (90) and (91) makes it possible to express the way in

which the amplification of a resistance-coupled amplifier varies with frequency by means of the universal amplification curve shown in Fig. 95. This gives the actual amplification relative to the amplification in the middle range of frequencies, as well as the relative phase shift, and is a universal curve that applies to all ordinary resistance-coupled amplifiers.

The procedure for obtaining the amplification characteristic by using the universal resonance curve (or Table VI) is: *first*, calculate the amplification in the middle-frequency range, using Eq. (88*a*) or (88*b*); *second*, calculate X_c/R for various frequencies in the low-frequency range (or, what is often more convenient, calculate the frequencies that give X_c/R

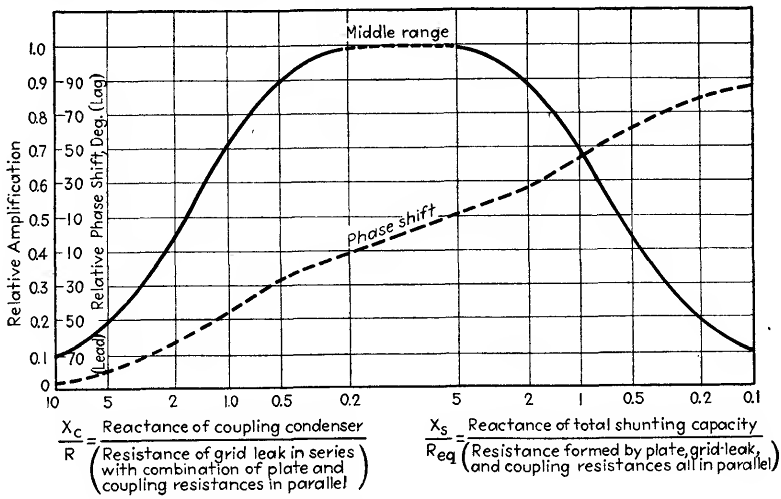


FIG. 95.—Universal amplification curve, showing how the amplification of a resistance-coupled amplifier falls off at high and low frequencies.

convenient values tabulated in Table VI), and then use the universal amplification curve (or Table VI) to get the low-frequency falling off; *finally*, estimate the total shunting capacity C_s and then calculate X_s/R_{eq} at various high frequencies (or, what is often more convenient, calculate the frequency at which X_s/R_{eq} has convenient values tabulated in Table VI), and then get the high-frequency response with the aid of the universal amplification curve (or Table VI).

Example.—If the above procedure is applied to the circuit constants of Fig. 93, the amplification in the middle range of frequencies is found by substitution in Eqs. (88*b*) and (89) to be

$$R_{eq} = \frac{500{,}000}{1 + \dfrac{500{,}000}{500{,}000} + \dfrac{500{,}000}{5{,}000{,}000}} = 238{,}000$$

$$\left.\begin{array}{l}\text{Amplification in}\\ \text{middle range}\end{array}\right\} = G_m R_{eq} = 414 \times 10^{-6} \times 238{,}000 = 98.4$$

At low frequencies, referring to Eq. (91), the amplification falls to 70.7 per cent of the middle-range value when the frequency is such as to make the reactance of the coupling condenser equal to $R = R_{gl} + \dfrac{R_c}{1 + \dfrac{R_c}{R_p}} = 955{,}000$ ohms. This is when $f = 1/2\pi CR = 33.4$ cycles. From Table VI one finds that at frequencies of 66.8, 16.7, and 6.7 cycles (which are, respectively, 2, 0.5, and 0.2 times 33.4 cycles) the amplification is 0.895, 0.447, and 0.196 times the middle-frequency range value of 98.4, or 88.1, 43.9 and 19.3 times, respectively. At high frequencies the amplification falls to 70.7 per cent of its middle-range value when the reactance of the shunting capacity C_s equals R_{eq} [see Eqs. (89) and (90)]. This is at a frequency f such that

$$f = \frac{1}{2\pi C_s R_{eq}} = \frac{1}{2\pi \times 70 \times 10^{-12} \times 238{,}000} = 9{,}560 \text{ cycles}$$

By use of Table VI it is found that at frequencies of 4,780, 19,120, and 47,800 cycles the amplification is 0.895, 0.447, and 0.196 times the middle-frequency range value of 98.4, or 88.1, 43.9 and 19.3 times, respectively.

Factors Affecting the Design of Resistance-coupled Amplifiers.—The first consideration involved in the design of a resistance-coupled amplifier is the selection of a suitable tube. Pentodes such as used in radio-frequency amplifiers, and having a sharp cut-off (not of the variable-mu type), are generally favored because they give more amplification than triodes and also have a better frequency response at the high frequencies. Triodes are used to some extent, with high-mu tubes being preferred. This is because examination of the equivalent circuits of Fig. 92 shows that the amplification obtained with resistance coupling cannot exceed the amplification factor.

The choice of the coupling resistance in audio-frequency amplifiers is normally made on the basis of obtaining the maximum possible amplification, although it is also often necessary to keep in mind the amount of output voltage obtainable without excessive distortion. If other things were equal, the higher the coupling resistance, the greater would be the amplification; but under normal circumstances other things are not equal. Ordinarily there is available a fixed plate-supply voltage, and the problem is to make the best possible use of this potential. The greater the coupling resistance under such circumstances, the greater will be the voltage drop in this resistance, and hence the less will be the voltage actually left to apply to the plate of the tube. In the case of triodes this means that, as the coupling resistance is made greater, the plate current decreases, causing the plate resistance of the tube to increase, and thereby counteracting, at least partially, the benefits that would otherwise result from the greater coupling resistance. With pentodes, increasing the coupling resistance without changing the control-grid and screen potentials may reduce the plate voltage to the point where a virtual cathode

forms in front of the suppressor grid. The tube then fails to operate in the desired manner. *It is hence necessary with pentode tubes to reduce the space current as the coupling resistance is increased.* This reduces the mutual conductance of the tube, and therefore tends to counteract the increased amplification that would otherwise result from the greater coupling resistance. As a result of this situation, the amplification is not particularly critical with respect to coupling resistance.

In audio-frequency amplifiers employing pentode tubes it is customary to operate the tube under conditions such that the voltage at the plate is either about 100 volts or about 25 to 50 per cent of the supply voltage, whichever gives the lower voltage at the plate. This ordinarily leads to coupling resistances of the order of 100,000 to 500,000 ohms for pentodes. With triodes, the coupling resistance will depend upon the amplification factor of the tube. With high-mu tubes it will commonly be approximately equal to the plate resistance, and will ordinarily consume from one-fourth to one-half of the available plate-supply voltage, whereas with moderate-mu tubes the coupling resistance will be commonly two to four times the plate resistance. These proportions for pentodes and triodes correspond to a good compromise between conditions that give maximum amplification and conditions that are favorable for maximum undistorted output voltage.

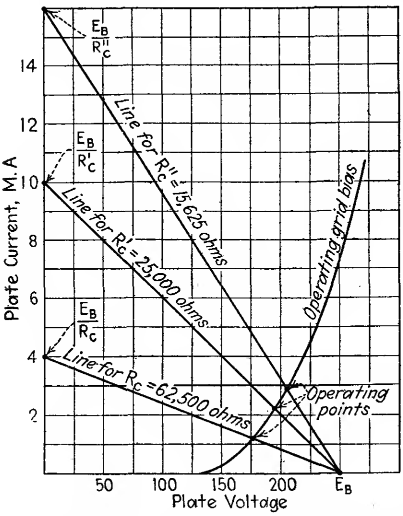


FIG. 96.—Graphical construction for determining the operating point of a tube when the plate-supply voltage E_B and the load resistance R_c are known.

The determination of the exact operating point (*i.e.*, plate current and plate voltage) of a triode resistance-coupled amplifier is complicated by the fact that the plate current depends on the voltage at the plate, while at the same time the voltage at the plate depends on the voltage drop produced by the plate current in the coupling resistance. It is possible to determine the plate potential for any given plate-supply voltage and coupling resistance by the trial-and-error process of assuming various plate currents and calculating the resulting plate voltage until the plate voltage that will give the assumed plate current is obtained. An alternative procedure is the geometrical construction shown in Fig. 96. Here one starts with a set of plate-current plate-voltage characteristic curves of the tube and finds the operating point as the intersection of the appro-

priate characteristic curve with a straight line that is determined by the load resistance as shown in Fig. 96. This line is drawn so that it intersects the horizontal axis at a plate potential equal to the plate-supply voltage and intersects the vertical axis at a current equal to the plate-supply voltage divided by the coupling resistance. Several such lines are shown in Fig. 96.

The control grid must be biased sufficiently negative to prevent the grid from drawing current at all times. About 1 volt negative bias is required to overcome the velocity of emission and contact potentials, so that the actual bias must be at least 1 volt more than the crest amplitude of the signal.[1] With triodes the bias should be no greater than necessary, and values of -1.3 to -1.5 volts are common for high-mu tubes. With pentode tubes somewhat higher values of bias, commonly -2 to -3 volts, are employed and the screen is then given the proper potential to draw the desired plate current with the bias used. The bias is commonly obtained as shown in Fig. 89 by the use of a resistance between cathode and ground having a value that gives the desired drop when the total space current flows through it. This bias resistor must be by-passed with a condenser large enough to have a reactance somewhat less than the bias resistance at the lowest frequency that is to be amplified reasonably well.

The grid-leak resistance and coupling-condenser capacity are selected on the basis of the desired low-frequency response, since from Eq. (91) it is seen that the smaller the ratio X_c/R (or, what is the same thing, the larger the product RC_c), the better will be the response at low frequencies. As already pointed out, the amplification falls to 70.7 per cent of its middle-range value when $R = X_c$. If this is to take place at a frequency f_0, then one must have $RC_c = 1/2\pi f_0$, where R is in megohms and C in microfarads. While there is an infinite number of resistance and capacity combinations that will make RC_c have a particular value, the most satisfactory proportion is when the grid-leak resistance is as high as possible, since then the shunting effect of the grid-leak resistance upon the coupling resistance in the middle range of frequencies is smallest and the amplification is consequently high [see Eq. (88*b*)]. The maximum value of grid-leak resistance that can be used is limited, however, by the fact that under ordinary circumstances the output of a resistance-coupled amplifier is applied to another tube. The grid-leak resistance is then in series with the grid of this tube, and, as explained in Sec. 38, must not be too high if possible damage to the output tube as a result of traces of gas is

[1] The contact potential varies from tube to tube, and also changes during the life of any one tube. In the case of high-mu triodes these changes, even if only a few tenths of a volt, appreciably alter the plate current and the mutual conductance at the operating point. This is one of the reasons pentodes are more satisfactory for resistance-coupled amplification than high-mu triodes.

to be avoided. Permissible values of grid-leak resistance for different types of tubes are given in the tube manuals.[1]

The coupling condenser must have a high leakage resistance. This is because the plate voltage of the tube is applied to the coupling condenser, and, if there is any leakage current, this will flow through the grid-leak resistance and produce a voltage across the grid leak which tends to bias the grid of the output tube positively. This effect can be counteracted by an additional negative bias, but such a procedure is not desirable because of the erratic behavior of the leakage currents through a condenser. Since the leakage resistance is inversely proportional to condenser capacity, changing the size of condenser while keeping the product RC_c constant has no effect upon the voltage that the leakage currents develop across the grid leak. The practical ultimate limit of low-frequency response is hence determined by the leakage of the coupling condenser. Mica coupling condensers will permit extending the low-frequency response to below 1 cycle, while good paper condensers are satisfactory for frequencies below any of those important in audio-frequency work. Poor paper condensers should never be used.

While it is a simple matter to maintain the full response to very low frequencies before appreciable falling off occurs, it is generally undesirable to provide a better low-frequency characteristic than is absolutely necessary. This is because such improvement of the low-frequency response is of no benefit, and at the same time invites trouble from feed-back effects arising from common plate impedances (see Sec. 51).

The response at high frequencies is determined by the shunting capacity and by the plate, coupling, and grid-leak resistances; it will be better the smaller the capacity and the lower the resistances. The shunting capacity is composed of the plate-cathode capacity (or output capacity) of the amplifier tube, a stray wiring capacity that is commonly from 4 to 10 $\mu\mu$f if the wiring is carefully arranged but which may be much greater if long leads are used, and the input capacity of the tube to which the amplified voltage is delivered. When the voltage is delivered to a pentode or screen-grid tube, the input capacity will be the sum of the grid-cathode and grid-screen tube capacities, and will commonly be 2 to 10 $\mu\mu$f. When the output voltage is delivered to a triode tube, the input capacity of this tube will consist of the grid-cathode capacity plus an additional capacity that takes into account the effect of the grid-plate

[1] With all small tubes, other than power tubes, the permissible grid resistance is usually of the order 1 megohm. With power tubes the maximum allowable resistance varies with different types of tubes, and also depends on whether or not self-bias is used. With self-bias it ranges from 100,000 for the 2A3 tube to 1,000,000 ohms for the 45 tube. With fixed bias the maximum permissible grid-leak resistance is so low that resistance coupling to power tubes with fixed bias is usually not recommended.

capacity of the triode.[1] This additional capacity depends upon the grid-plate capacity of the triode, upon the amplification factor, and upon the load impedance in the plate circuit, as explained in Sec. 50, and is usually much larger than all the other components of the shunting capacity put together.

In amplifiers intended for handling the audio-frequency range, the shunting capacity obtained with a properly arranged circuit is ordinarily so small that the coupling and other resistances can be chosen on the basis of values that will give the maximum possible amplification.[2] The response will then generally be satisfactory up to the highest audible frequencies without giving it any special consideration. If the high-frequency range is to be extended above the upper audible limit, as is necessary, for example, in television circuits, the equivalent resistance formed by the plate, coupling, and grid-leak resistance must be reduced. In the case of pentode tubes, which are normally used when high frequencies are involved, this means that the coupling resistance must be lowered. In the case of triodes, a tube with a lower plate resistance and hence a lower amplification factor is required. In either case the extension of the frequency range is accomplished at a sacrifice in amplification, which is the price that must be paid for the improved response.

Curves illustrating the characteristics of a resistance-coupled amplifier employing a pentode tube under a variety of typical conditions are given in Figs. 97, 98, 99, and 100. In Fig. 97 is shown the effect upon the low-frequency response of varying the coupling capacity C_c. Figure 98 shows how the high-frequency response becomes better as the shunting capacity becomes less. The effect of varying the coupling resistance, while leaving the other circuit elements the same, is shown in Fig. 99. In this figure the combination of control-grid and screen-grid potentials is changed as the coupling resistance is changed, in order to give a higher plate current as the coupling resistance becomes lower. It will be noted that the coupling resistance has very little effect on the relative falling

[1] To a first approximation the input capacity of an amplifier tube is given by the equation

$$\text{Input capacity} = C_{gf} + C_{gp}(1 + A) \tag{92}$$

where C_{gf} is the grid-cathode capacity, C_{gp} is the grid-plate capacity, and A is the ratio of the alternating voltage across the load impedance in the plate circuit of the tube to the signal voltage applied to the control grid. A cannot exceed the amplification factor μ of the tube and will normally be at least one-half the amplification factor. Equation (92) is a simple form of Eq. (114) for the special case where the load is a resistance.

[2] The only exception to this is when the output of the resistance-coupled amplifier is applied to a relatively large power tube, such as a Type 845 tube. In such a case the input capacity of the power tube is often large enough to spoil the response at the higher audio frequencies unless a low coupling resistance is used.

off at low frequencies, but that the *relative* falling off at high frequencies is less the lower the resistance, although the low resistance involves a sacrifice in amplification. The increased plate current, and hence increased mutual conductance, which go with a higher plate-supply voltage, are seen from Fig. 100 to produce an appreciable increase in the amplification.

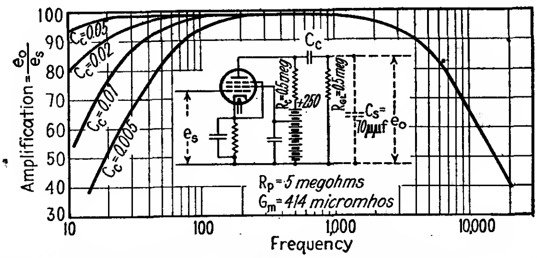


FIG. 97.—Effect of varying the coupling-condenser capacity C_c on the characteristics of a typical resistance-coupled amplifier.

Design Procedure and Typical Designs.—The steps that should be followed in designing a resistance-coupled amplifier based upon the above considerations will now be enumerated. In the first place, one must ascertain the plate-supply voltage that will be available and the frequency range required. The next step is to select a tube, which will normally be one of the pentode tubes having a sharp cut-off (not variable-mu) used for radio-frequency amplification. Under special conditions one may use a triode, generally a high-mu triode, but the pentode ordinarily gives

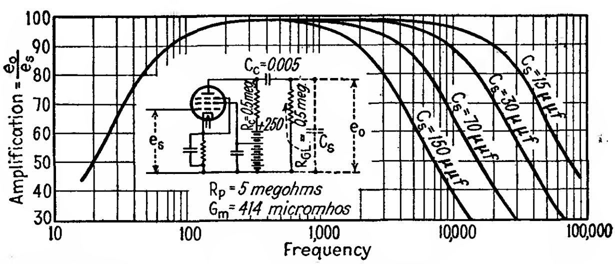


FIG. 98.—Effect of varying the shunting capacity C_s on the characteristics of a typical resistance-coupled amplifier.

a superior performance. The next step is the determination of the proper coupling resistance. This will depend upon the plate-supply voltage available and the frequency range desired. When pentode tubes are used to amplify the audio-frequency range, the coupling resistance will commonly lie between 100,000 and 500,000 ohms. With triode tubes the value will depend upon the amplification factor of the tube. With the coupling resistance tentatively selected, control-grid and screen-grid

potentials that appear to be satisfactory are chosen, after which the plate current, the voltage at the plate, and the mutual conductance are determined for the resulting operating point. It is then possible to determine the bias resistance required to give the assumed bias, and also the con-

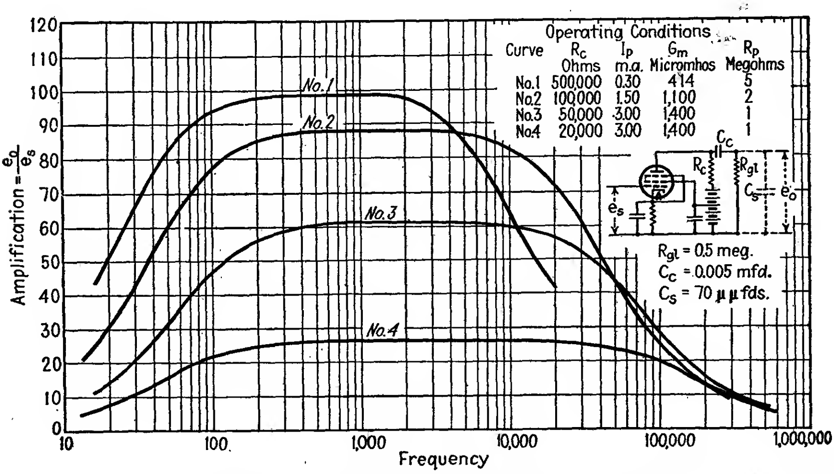


FIG. 99.—Effect that designing a typical resistance-coupled amplifier for different coupling resistances has on the amplification characteristic.

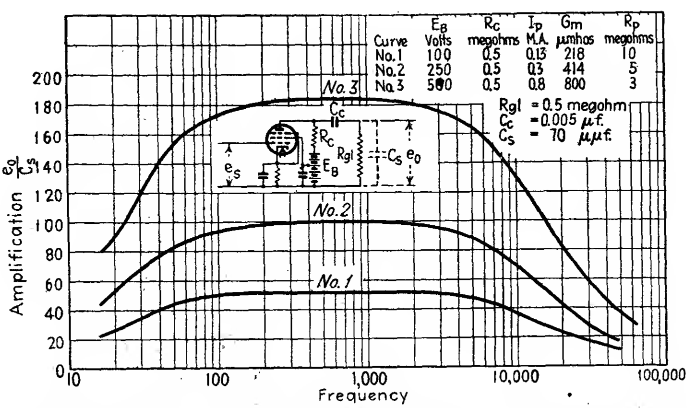


FIG. 100.—Effects produced by designing a typical resistance-coupled amplifier for different plate-supply voltages.

denser that will be required to by-pass this resistance. The next step is to choose as the grid leak the highest resistance that can be safely inserted in the grid circuit of the tube to which the amplifier output is to be applied. A coupling condenser is then used which will make the amplification hold up reasonably well to the lowest frequency that is to be amplified, but which will allow a falling off at lower frequencies.

Finally the shunting capacities are estimated, and the high-frequency response is calculated to make sure that it will be satisfactory.

In the case of resistance-coupled amplifiers which are required to amplify frequencies above the audio range, the above procedure is modified by estimating the shunting capacities before the coupling resistance is selected and then selecting the highest coupling resistance that will give the required high-frequency characteristic with this capacity.

Several typical designs of resistance-coupled amplifiers suitable for audio-frequency work are given in Table VII; they indicate what are reasonable proportions when employing pentodes or high-mu triodes.

TABLE VII.—TYPICAL DESIGNS OF RESISTANCE-COUPLED AMPLIFIERS[1]

	Pentode 57,77, and 6C6 tubes						Triode section of 2A6 and 75 tubes					
Plate supply, volts	100		180		250		100		180		250	
Screen supply, volts	20	20	30	30	50	52						
Control-grid bias, volts	−1.05	−1.25	−1.30	−1.55	−2.1	−2.3	−1.10	−1.05	−1.30	−1.30	−1.35	−1.35
Cathode resistor, ohms	3,400	7,250	2,600	4,850	3,500	6,200	11,550	15,000	5,450	9,000	3,380	5,600
Plate resistor, megohms	0.25	0.50	0.25	0.50	0.25	0.50	0.25	0.50	0.25	0.50	0.25	0.50
Grid resistor, megohms	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Plate current, milliamperes	0.23	0.13	0.38	0.24	0.48	0.30	0.09	0.07	0.24	0.14	0.40	0.24
Plate voltage as per cent of plate supply	42.5	35.0	47.2	33.3	52.0	40.0	77.5	65.0	66.6	61.1	60.0	52.0
Voltage amplification	54	53	92	93	100	110	36	37	56	55	59	58
Output voltage (peak volts) with only little distortion	22	23	44	48	65	65	17	17	36	34	41	40

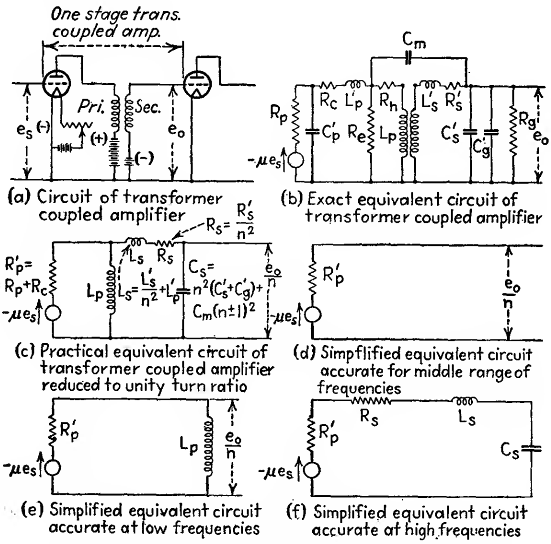
[1] These designs represent recommendations of RCA Radiotron Company.

45. Audio-frequency Voltage Amplifiers. Transformer Coupling.—In the transformer-coupled amplifier the load impedance connected in the plate circuit of the vacuum tube is the primary of a transformer, the secondary voltage of which is applied to the grid of the succeeding tube, as shown in Fig. 101*a*. Transformer-coupled amplifiers are used primarily for supplying the exciting voltage of power tubes. Their principal advantages compared with resistance coupling for this purpose are that the direct-current resistance connected in series with the grid of the tube to which the amplified voltage is applied is small, and that the maximum amount of undistorted voltage that can be developed is considerably higher in proportion to the plate-supply voltage than with resistance coupling. The transformer-coupled arrangement also adapts itself much more readily to the excitation of push-pull power amplifiers than does resistance coupling. The principal disadvantages of transformer coupling are the much higher cost as compared with resistance coupling and the fact that stray magnetic fields tend to induce voltages in the transformer.

Transformer coupling is generally used in connection with triode tubes as shown in Fig. 101. Pentode tubes can be used with transformer coup-

ling by connecting a resistance across the transformer as shown in Fig. 109, but they are not generally recommended because the triode tube is less expensive and is also capable of giving more undistorted output.

The most important characteristics of a transformer-coupled amplifier are the amount of amplification obtained and the way in which this amplification varies with frequency. A typical amplification curve is



Notation

R_p = Plate resistance of tube
R_c = D.C. resistance of primary
L_p' = Primary leakage inductance
R_e = Resistance representing eddy current loss
R_h = Resistance representing hysteresis loss
L_p = Incremental primary inductance
R_s' = Resistance of secondary winding
L_s' = Secondary leakage inductance
C_s' = Secondary distributed capacity
C_p' = Capacity in shunt with primary (tube plus transformer capacity)
C_m = Capacity between primary and secondary windings of the transformer
C_g' = Input capacity of tube to which voltage e_o is delivered
R_g = Input resistance of tube to which voltage r_1 is delivered
n = Ratio of secondary to primary turns
μ = Amplification factor of tube

FIG. 101.—Actual circuit of a transformer-coupled amplifier, together with exact and approximate equivalent circuits.

shown in Fig. 102. The distinguishing features of this curve are relatively constant amplification in the middle range of frequencies, but with a gradual falling off at low frequencies and a much sharper falling off at high frequencies. The loss in amplification at low frequencies results from the low reactance that the transformer primary has at low frequencies, while the falling off at high frequencies is a result of the leakage inductance and distributed capacity of the transformer. The frequency limits of the amplifier depend largely upon the design of the transformer, and can be readily made to cover the voice-frequency range.

Difficulties are encountered in amplifying very low and very high frequencies, however, and resistance coupling is usually preferred where very wide bands of frequencies are involved.

Analysis of Transformer-coupled Amplifiers Used with Triode Tubes.—In analyzing the behavior of a transformer-coupled amplifier, the proper procedure is first to draw an equivalent circuit in which the tube is replaced by its equivalent electrical circuit. The result for the case of a triode tube, taking into account all essential features of the transformer, is shown in Fig. 101*b*. The only approximation involved in this circuit is in assuming that the coil capacities can be represented as lumped when

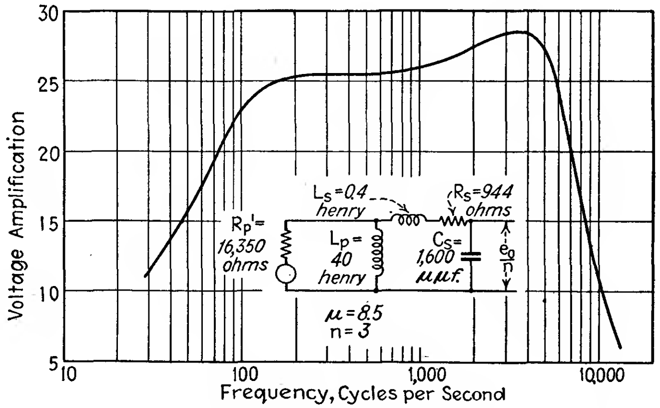


FIG. 102.—Variation of amplification with frequency in a typical transformer-coupled amplifier.

they are distributed in fact. The transformer is seen to be a very complicated electrical network, the solution of which involves a great deal of labor, but fortunately it is possible to reduce this exact circuit into the very much simpler network shown at Fig. 101*c* without introducing appreciable errors. The transition from the exact circuit of Fig. 101*b* to the practical circuit shown at Fig. 101*c* consists in neglecting the effects of the distributed capacity of the primary and the plate-cathode capacity of the amplifier tube, in neglecting the resistances resulting from eddy currents and magnetic hysteresis, in reducing the transformer to a unity-turn ratio, in lumping all the leakage inductance on the secondary side of the transformer, and in considering the primary to secondary (*i.e.*, interwinding) capacity C_m as equivalent to a certain other capacity shunted across the secondary of the transformer.[1]

[1] These approximations can be justified as follows: The primary capacity C_p' and the resistance R_h resulting from hysteresis have negligible effect on the behavior of the transformer because of their small size in ordinary circumstances. The reduction to a unity-turn ratio can always be carried out without introducing error

The resistance R_e representing the losses from eddy currents is usually so high as to have negligible shunting effect, and so need not be considered unless extreme accuracy is required. The effect of R_e can be readily taken into account by using Thévenin's theorem to combine the resistance R_e with the plate resistance. Doing this shows that the effect of the eddy-current resistance R_e can be taken into account by using slightly reduced values of R_p and μ in Fig. 101c and Eqs. (94), (95), and (96), according to the relations:

$$\left.\begin{matrix}\text{Effective } \mu \text{ when taking}\\ \text{into account } R_e\end{matrix}\right\} = \frac{\text{actual } \mu \text{ of tube}}{1 + \dfrac{R_p}{R_e}} \tag{93a}$$

$$\left.\begin{matrix}\text{Effective } R_p \text{ when taking}\\ \text{into account } R_e\end{matrix}\right\} = \frac{\text{actual value of } R_p}{1 + \dfrac{R_p}{R_e}} \tag{93b}$$

where R_p is defined as in Fig. 101. When the proper values are assigned to the circuit elements in Fig. 101c, this simplified equivalent circuit of the transformer-coupled amplifier will give calculated results that are in error by only a few per cent.

The complexity of the equivalent circuit of the transformer-coupled amplifier can be greatly reduced if limited frequency ranges are considered at a time. Thus in the middle range of frequencies the reactance of the transformer primary inductance is so high as to be essentially an open circuit, while at the same time the small distributed and shunting capacity C_s is substantially an open circuit. With these simplifications the equivalent circuit reduces to that shown in Fig. 101d, and the amplification then becomes the product of turn ratio and amplification factor. That is,

$$\left.\begin{matrix}\text{Amplification in middle}\\ \text{range of frequencies}\end{matrix}\right\} = \frac{e_0}{e_s} = \mu n \tag{94}$$

Somewhere in this middle range of frequencies the shunting capacity C_s is in parallel resonance with the primary inductance L_p, causing the

provided all impedances on the secondary side of the transformer are divided by the square of the voltage step-up ratio when placed in the equivalent circuit. The leakage inductance can be considered as located entirely on the secondary side because this is where most of it actually is and because at the high frequencies where leakage inductance becomes important its exact distribution between primary and secondary is not important. The interwinding capacity C_m can be replaced by a suitable capacity located across the secondary as a result of the fact that the voltage difference across this condenser is almost exactly in phase with the voltage developed across the secondary of the transformer up to the frequency at which the secondary capacity is resonant with the leakage inductance; furthermore this voltage is proportional to the voltage across the secondary but has a magnitude either $(n + 1)$ or $(n - 1)$ (depending on the relative polarity of the two windings) times the voltage across the primary of the transformer.

impedance across the primary terminals of the transformer to be extremely high.

At low frequencies the shunting capacity is still less important than in the middle range of frequencies, but the reactance of the primary inductance of the transformer tends to become low and so must be taken into account. This gives the equivalent circuit of Fig. 101e, which is applicable to low frequencies and from which a simple analysis shows:

$$\left.\begin{matrix}\text{Amplification at low}\\ \text{frequencies}\end{matrix}\right\} = \frac{e_0}{e_s} = \mu n \frac{1}{[1 + (R_p'/\omega L_p)^2]^{1/2}} \tag{95}$$

Here $\omega L_p/R_p'$ is the ratio of the reactance of the primary inductance to the resistance formed by the plate resistance of the tube plus the direct-current resistance of the primary winding. It will be noted that the extent to which the amplification at low frequencies falls below the amplification μn obtained in the middle-frequency range depends only upon the ratio $\omega L_p/R_p'$, and in general drops off with frequency according to the same general law as does the low-frequency amplification of a resistance-coupled amplifier. The extent of the falling off can be estimated from the fact that, when the frequency is such that ωL_p equals R_p', the amplification has dropped to 70.7 of its mid-range value.

The primary inductance that is effective in the equivalent circuits of Fig. 101 and in Eq. (95) is the inductance to the alternating component of the plate current that is superimposed upon the direct-current plate current of the tube. This inductance is hence determined by the incremental permeability that was discussed in Sec. 6, and will be greater as the direct-current magnetization decreases and as the alternating flux density increases.

At high frequencies the reactance of the primary inductance is so high as to be an entirely negligible shunt, but the shunting capacity C_s and the leakage inductance can no longer be ignored. This results in the equivalent circuit of Fig. 101f, which is a series-resonant circuit having a very low Q. A simple calculation based upon this circuit shows that

$$\left.\begin{matrix}\text{Amplification at}\\ \text{high frequencies}\end{matrix}\right\} = \frac{e_0}{e_s} = \mu n \frac{1/\omega C_s}{\sqrt{(R_p' + R_s)^2 + \left(\omega L_s - \dfrac{1}{\omega C_s}\right)^2}} \tag{96a}$$

$$= \mu n \frac{1}{[(\gamma/Q_0)^2 + (\gamma^2 - 1)^2]^{1/2}} \tag{96b}$$

where

$$\gamma = \frac{\text{actual frequency}}{\text{series-resonant frequency of } C_s \text{ and } L_s}$$

$Q_0 = \dfrac{\omega_0 L_s}{R_p' + R_s}$ = circuit Q at the frequency for which C_s and L_s are in series resonance.

These equations show that the amplification at high frequencies depends both upon the plate resistance of the tube and upon the series resonance action between the leakage inductance and secondary capacity. For the best characteristic at high frequencies a careful balance must be realized, as discussed below.

Universal-amplification Curve of Triode Transformer-coupled Amplifier.—The simplicity of Eqs. (94), (95), and (96*b*) makes it possible to express the relative amplification of a transformer-coupled amplifier in the universal amplification curve of Fig. 103. This gives the relative ampli-

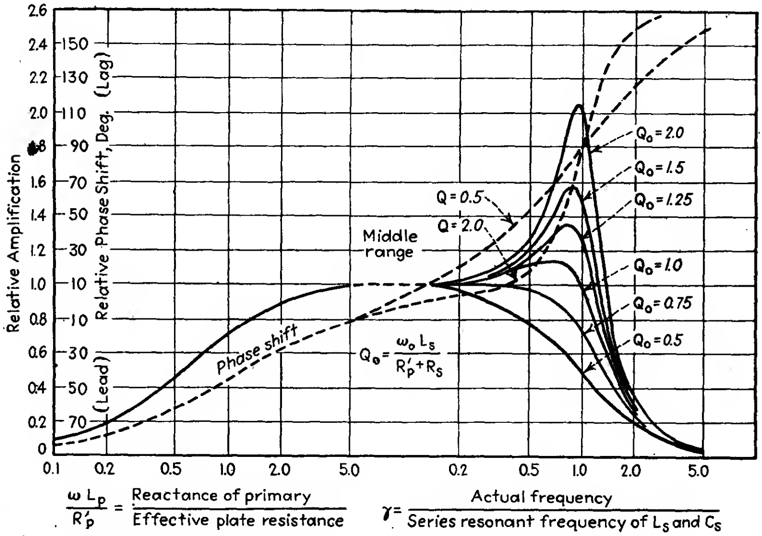


FIG. 103.—Universal amplification curve of transformer-coupled amplifiers.

fication and phase shift as a function of frequency, and applies to all transformers. The actual amplification at any frequency is the relative amplification obtained from the universal resonance curve of Fig. 103 multiplied by the quantity μn.

The low-frequency part of the universal amplification curve is a plot of Eq. (95) showing the way in which the amplification falls off as a function of $\omega L_p/R_p'$. The frequency factor is introduced through its effect on the primary reactance ωL_p. The high-frequency portion of the universal amplification curve is a plot of Eq. (96*b*) showing the variation of amplification as a function of the quantity γ for various values of Q_0. The frequency is introduced in the high-frequency range through its effect upon γ.

To use the universal amplification curve of Fig. 103, one first calculates the amplification in the middle range according to Eq. (94). The next step is to determine the frequency at which $\omega L_p/R_p' = 1$, and from

this the frequency at which $\omega L_p/R_p'$ has other convenient values such as 2, 0.5, 0.2, etc. The falling off in amplification at these various frequencies is then obtained directly from the low-frequency portion of the universal amplification curve. Finally the frequency at which the leakage inductance L_s and the shunting capacity C_s are in series resonance is calculated; and also the value of the circuit Q at this frequency, *i.e.*, Q_0. This frequency corresponds to a value $\gamma = 1$. The high-frequency response can then be obtained readily from the high-frequency portion of the universal amplification curve for other frequencies simply related to the series-resonant frequency, such as 0.5, 2, 5 times the series-resonant frequency.

Discussion of Amplification Characteristic.—A consideration of the factors involved in transformer-coupled amplification shows that these are relatively simple in spite of the complexity of the equivalent circuit of the transformer. At low frequencies the falling off in amplification is controlled by the ratio $\omega L_p/R_p'$, which means that, in order to obtain a good low-frequency response, the incremental primary inductance of the transformer must be high and the plate resistance of the tube low. Hence for good low-frequency response the tube should not have too high an amplification factor, since the higher the amplification factor, the higher will normally be plate resistance. The primary winding of the transformer should also have a large number of turns, which with a fixed winding space leaves less space for the secondary, and hence results in a low step-up ratio and low amplification.

Passing now to the characteristics at high frequencies, examination of Fig. 103 shows that the best response is obtained when the value of Q_0 is equal to or slightly less than unity, and that when this condition exists the amplification is substantially constant up to the series-resonant frequency of L_s and C_s, but falls off rapidly at higher frequencies. This means that for any particular transformer there is a proper value of plate resistance that should be used. If the plate resistance is lower than this optimum value, the response at low frequencies will be improved, but there will be a peak at the high frequencies. Conversely, if the plate resistance is much greater than the optimum value, both high and low frequencies will be adversely affected.

It is apparent from the above discussion that the tube and transformer must be designed to cooperate with each other to produce the desired result. Most audio-frequency transformers are designed to operate in the plate circuits of triodes having plate currents of 2 to 8 ma and plate resistances of the order of 8000 to 15,000 ohms at the operating point. These conditions correspond to a general-purpose tube having an amplification factor of 8 to 20, and represent the situation that has been found to be most satisfactory with transformer coupling. The plate current is

low enough to prevent excessive direct-current saturation in the transformer primary, the plate resistance is low enough to insure the possibility of a satisfactory low-frequency response, and large output voltages can be developed without distortion.

The grid bias should be so chosen in relation to the available plate-supply voltage as to give the plate resistance with which the transformer should operate. Adjustments of the grid bias will enable the high-frequency response to be changed appreciably. In addition, the plate-supply voltage should be great enough so that, when the proper plate resistance is realized, the bias will be sufficient to maintain the grid negative at the crest of the largest signal voltage to be amplified.

Transformer Characteristics.—From the above considerations it is seen that the ideal transformer has a high turn ratio, a high primary inductance, a very low leakage inductance, and the highest possible series resonant frequency (*i.e.*, low shunting capacity as well as low leakage inductance). In attempting to approach this ideal, a number of practical conflicts are encountered. With a given-size core, the total number of turns is fixed; so, if a high turn ratio is used, only a few of the turns can be allotted to the primary, which results in a low primary inductance. Increasing the size of the core and winding space to give a satisfactory primary inductance with a high turn ratio will not remedy the situation since the leakage inductance will be increased as a result of the larger size, and this reduces the series resonant frequency and hence spoils the high-frequency response. The leakage inductance can be reduced greatly by interleaving the primary and secondary windings, but this increases the distributed capacity and so tends to neutralize the benefits obtained from the lower leakage inductance.

In order to have a good response, a transformer must have a large core of high-quality magnetic material, with an air gap in the core sufficient to prevent the normal direct-current plate current of the amplifier tube from producing undue saturation, and yet not so large as to reduce the incremental inductance of the primary winding. The step-up ratio n should be reasonably low, commonly 2 to 4, and improved results can be obtained by a certain amount of interleaving of primary and secondary windings. The use of a large core of high-grade magnetic material with proper air gap provides a minimum of reluctance in the magnetic circuit. The low turn ratio makes it possible to obtain a relatively high primary inductance and also tends to improve the high-frequency response by increasing the series-resonant frequency of the transformer. The interleaving when properly done will also increase the series resonant frequency.

If the secondary capacity C_s is to be kept small, it is important that the secondary terminal leading to the grid of the output tube be from the

outside of the secondary winding. The proper arrangement is normally marked on the transformer. The arrangement of the primary terminals is relatively unimportant.

The quality of the magnetic material in the core is of particular importance and, if unusually severe requirements must be met, high-permeability alloys such as permalloy should be used. The reason for this is that the primary inductance of a transformer is determined by the reluctance of the magnetic circuit provided by the core, while the leakage inductance and the series resonant frequency are caused by flux paths in air and so are independent of the core. Improvements in core material will therefore increase the primary inductance in proportion to leakage

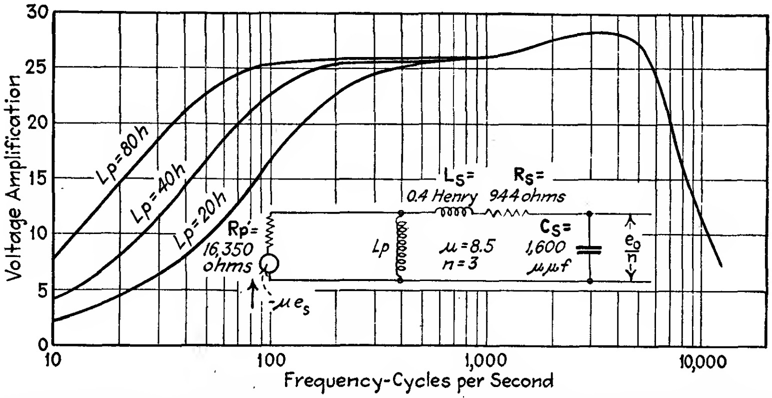


FIG. 104.—Effects that would be produced on the frequency response of a transformer-coupled amplifier by varying the primary inductance. Note the improved low-frequency characteristic when the primary inductance is large.

inductance and distributed capacity. This permits a wider frequency range than would be possible with a transformer having the same bulk but using poorer magnetic material, or, conversely, for the same frequency range means that high-permeability core material makes the transformer smaller and lighter. When using magnetic alloys having high permeability, it is necessary to employ a shunt-feed arrangement (see next section) since the incremental inductance of high-permeability alloys is very adversely affected by direct-current magnetization.

It is possible to determine the characteristics of an amplifying transformer with fair accuracy by calculation based upon design data.[1] While such calculations are empirical and so depend to a considerable extent upon the experience and judgment of the computer, they are extremely useful in aiding the designer of audio-frequency amplifier transformers to improve his product. By properly manipulating the factors, such as

[1] See Glenn Koehler, The Design of Transformers for Audio-frequency Amplifiers with Preassigned Characteristics, *Proc. I.R.E.*, vol. 16, p. 1742, December, 1928.

core dimensions, shape, space, and arrangement of windings, air gap in the magnetic circuit, etc., it is possible for a skillful designer to increase materially the frequency range of a transformer.

Typical Characteristics.—The effect that various circuit elements have upon the amplification characteristic of a transformer-coupled amplifier

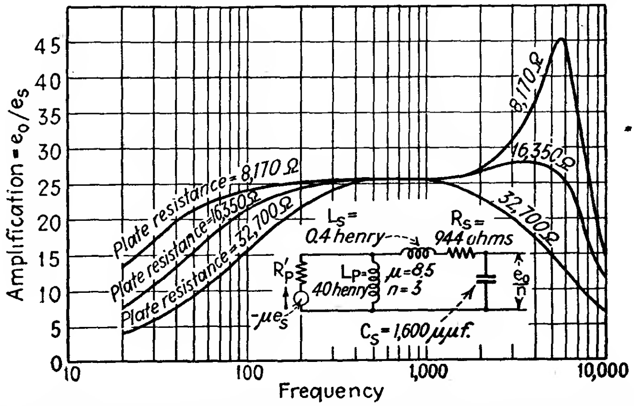


FIG. 105.—Effect of varying the plate resistance in a transformer-coupled amplifier. Note that the best characteristic is obtained when the plate resistance is neither too high nor too low, but rather is properly matched to the transformer.

are shown in Figs. 104 to 107. It will be observed that varying the primary inductance while leaving everything else the same affects only the low-frequency response. On the other hand, varying the effective plate resistance R_p' alters both the low- and high-frequency response, with a high plate resistance giving a falling off at both low and high

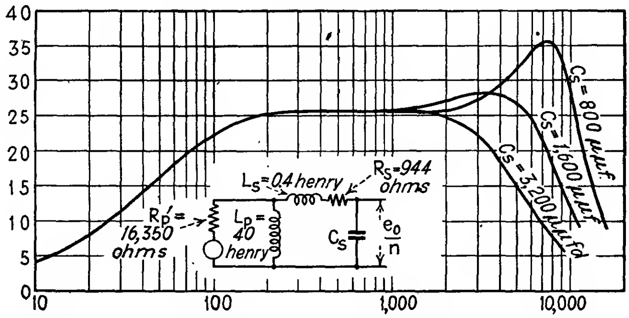


FIG. 106.—Effect of varying the effective secondary shunting capacity C_s in a transformer-coupled amplifier.

frequencies, while a low plate resistance improves the low-frequency response but introduces a peak at the high frequencies. Shunting capacity across the transformer secondary, or in some other way altering the distributed capacity of the secondary, is seen to affect only the high-

frequency response. Increasing the capacity reduces the series resonant frequency and also lowers the value of Q_0, thereby causing a pronounced falling off at higher frequencies, whereas reducing the distributed capacity extends the high-frequency range, although at the same time producing a high-frequency peak unless the plate resistance is increased. Varying the leakage inductance alters both the series-resonant frequency and the value of Q_0. It will be noted that some leakage inductance is desirable, since with no leakage whatsoever the shunting capacity causes a falling off of the high-frequency response much sooner than if the appropriate amount of leakage were present.

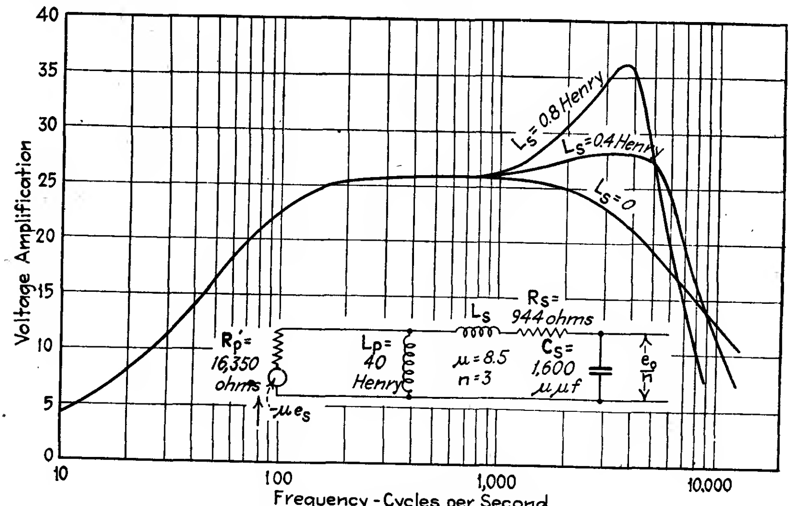


FIG. 107.—Effect of varying the leakage inductance of the secondary. Note that some leakage inductance is desirable in order to neutralize the detrimental effects of the effective capacity C_s shunted across the secondary.

Resistances are sometimes shunted across either the primary or secondary windings of a transformer in order to control the amplification characteristic. A resistance shunted across the primary has the effect of reducing the effective amplification factor μ and the effective plate resistance R_p.[1] This reduces the amplification, improves the low-frequency response, and tends to make the high-frequency response peaked.

When the transformer has a peak at high frequencies and a poor response at low frequencies, considerable improvement in the frequency

[1] The behavior of the arrangement can be readily analyzed by using Thévenin's theorem to simplify the equivalent network that supplies power to the primary terminals of the transformer, exactly as was done in Eqs. (93*a*) and (93*b*) to take into account the eddy-current resistance R_e that is shunted across the primary terminals. In fact, these same equations can be used to take into account the action of a shunting resistance by considering R_e to be the shunting resistance.

response can be obtained at the expense of amplification by shunting a resistance across the transformer secondary. Such a shunting resistance lowers the amplification but affects the high frequencies most and the low frequencies least so that, when the proper value of resistance is used, the response characteristic is more uniform, as illustrated in Fig. 108.[1]

Transformer-coupled Amplifier with Pentode Tubes.—Transformer coupling can be used with pentode tubes by shunting a suitable resistance across the primary of the transformer, as shown in Fig. 109*a*. *The proper value of resistance to connect across the primary terminals can be determined from the fact that the frequency distortion in a pentode transformer-coupled*

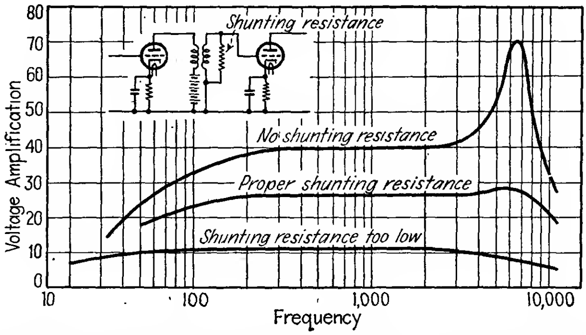


FIG. 108.—Effect of shunting resistances across the secondary of the transformer-coupled amplifier, showing how a transformer having a high-frequency peak can be improved by making the resistance the proper value, but at the expense of reduced amplification.

amplifier is exactly the same as the frequency distortion that would be present in a triode amplifier employing the same transformer and having a plate resistance equal to the equivalent resistance formed by the shunting resistance in parallel with the plate resistance of the pentode. Since the plate resistance of the pentode tube is very large, this means that the resistance shunted across the primary should approximate the plate resistance that would be used with the same transformer in a triode amplifier. This results from the fact that according to Thévenin's theorem the important thing as far as the transformer behavior is concerned is the resistance of the network that the transformer sees when looking toward the tube from its primary terminals. The equivalent circuit of the pentode transformer-coupled amplifier can accordingly be reduced by the use of Thévenin's theorem to the arrangement shown in Fig. 109*b*, which is identical with the equivalent circuit for the triode case. Hence a separate analysis of the pentode transformer-coupled amplifier is not required.

[1] More complete discussion of transformer-coupled amplifiers having resistance shunted across the secondary is given by Paul W. Klipsch, Design of Audio-frequency Amplifier Circuits Using Transformers, *Proc. I.R.E.*, vol. 24, p. 219, February, 1936.

The use of Thévenin's theorem also shows that the amplification obtained with both triodes and pentodes is proportional to the mutual conductance, provided the resistance that the transformer primary terminals see when looking toward the tube is the same for both types of tubes. Since the mutual conductance of pentodes and triodes is roughly the same for the same plate currents, there is no choice between the tubes from the point of view of amount of amplification obtainable with transformer coupling. At the same time, the pentode amplifier tends to have more amplitude distortion in proportion to its output voltage, requires a more expensive tube, and also demands means for supplying voltage to the screen as well as to the plate. For these reasons triode tubes are normally preferred for transformer-coupled amplifiers.

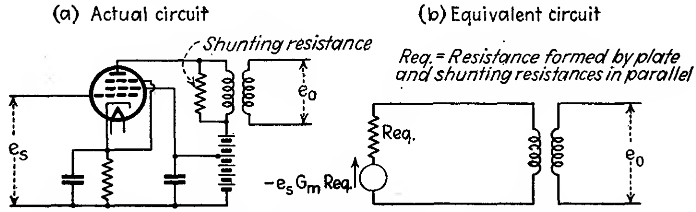


FIG. 109.—Actual and equivalent circuit of pentode transformer-coupled amplifier.

Experimental Determination of Transformer Characteristics.—The primary inductance L_p that is effective in the equivalent circuits of Fig. 101 is the incremental inductance measured with low alternating flux density in the presence of the appropriate direct current. The results obtained will depend somewhat upon the previous magnetic history of the core, and usually cannot be duplicated more closely than within a few per cent unless the core is first carefully demagnetized.[1]

The eddy-current resistance R_e can usually be ignored. It is to a high degree of accuracy equal to the impedance at the primary terminals when the frequency is such that the capacity C_s and the primary inductance L_p are in parallel resonance.

The step-up ratio of the transformer may be most satisfactorily measured by applying a known voltage to the primary and measuring the potential at the secondary terminals with a vacuum-tube voltmeter. This measurement must be carried out at a low frequency, *i.e.*, a few hundred cycles or less, since at higher frequencies incipient resonance between the capacity C_s and the leakage inductance L_s will make the voltage that appears at the secondary terminals higher than the voltage actually induced in the secondary winding. When measurements of step-up ratio at low frequencies are made, the actual voltage acting on

[1] For details as to incremental inductance measurements see the author's "Measurements in Radio Engineering," pp. 53–58, 1st ed. McGraw-Hill Book Company, Inc.

the primary inductance is less than the voltage applied to the primary terminals by the drop in the direct-current resistance of the primary winding. The error thus introduced can be allowed for by multiplying the apparent step-up ratio by the factor $\sqrt{1 + (R_c/\omega L_p)^2}$, where R_c and L_p are primary resistance and inductance, respectively.

The leakage inductance L_s reduced to unity turn ratio can be readily determined by measuring the inductance between the primary terminals when the secondary is short-circuited. The leakage inductance is independent of frequency and of saturation effects in the core, since the leakage paths are in air; thus the measurement may be made at any frequency and without passing direct current through the primary. It is not permissible to measure leakage inductance from the secondary terminals by short-circuiting the primary since errors are then introduced as a result of the secondary capacity C_s.

Several indirect methods can be used to obtain the effective secondary capacity C_s. One way is to determine the frequency $\omega_1/2\pi$ at which parallel resonance occurs across the secondary terminals when the primary terminals are short-circuited. This frequency is that at which C_s and L_c are in parallel resonance, so that

$$C_s = \frac{1}{\omega_1{}^2 L_s} \tag{97}$$

Another way of getting at the same thing is to measure the ratio of secondary to primary voltage as a function of frequency at frequencies approaching the resonant frequency. This ratio will increase with frequency and reach a maximum when L_s and C_s are in resonance. At some frequency below resonance, C_s can be deduced from a knowledge of L_s and the step-up ratio according to the relation

$$\frac{\text{Voltage ratio of transformer}}{\text{Step-up ratio of transformer}} = -\frac{1/\omega C_s}{\omega L_s - \dfrac{1}{\omega C_s}} \tag{98}$$

Other indirect methods of determining C_s can be devised when the occasion requires. In measuring C_s it must be kept in mind that the capacity that is effective is the sum of the distributed secondary capacity of the transformer and the input capacity of the tube to which the secondary delivers its voltage. Consequently C_s should be determined with the output tube actually present and operating with its normal load impedance. Any additional measuring equipment, such as vacuum-tube voltmeters, across the secondary must be so arranged as to introduce the minimum possible additional shunting capacity.

Push-pull Transformers.—One of the chief uses of transformer coupling is to excite the grids of a push-pull power amplifier. The circuit

arrangement involved is shown in Fig. 110 and requires a center-tapped secondary winding. When the two halves of the secondary are symmetrical with respect to the primary, as in Fig. 111*a*, the push-pull transformer behaves in exactly the same way as the ordinary transformer, which has already been discussed. The only difference, then, is that the amplified voltage applied to each of the output tubes is half of the total voltage developed by the full secondary winding. In analyzing the behavior of a symmetrical push-pull transformer, one uses the full secondary, and determines the equivalent leakage inductance, series-resonant frequency, etc., just as though the center-tap did not exist.

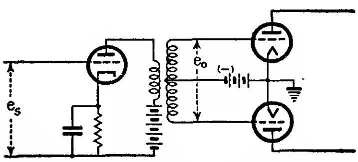


FIG. 110.—Circuit of push-pull transformer-coupled amplifier.

In the cheaper push-pull transformers, such as are commonly used in radio receivers, the usual method of construction is shown in Fig. 111*b*. With this arrangement the two sides of the secondary are unsymmetrically located with respect to the primary. The half of the secondary next to the core has a higher shunting capacity C_s than does the other half, but likewise possesses a lower leakage inductance with respect to the primary winding. As a consequence, the voltages delivered by the two halves of the secondary winding are not the same in the high-frequency range, although at medium and low frequencies the dissymmetry has no effect.

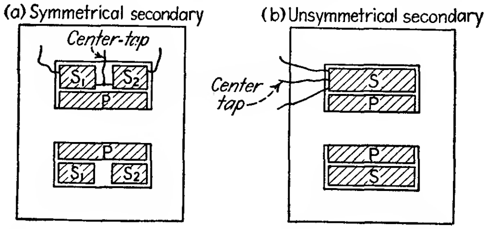


FIG. 111.—Typical winding arrangements in push-pull-amplifier transformers.

46. Audio-frequency Voltage Amplifiers. Miscellaneous Coupling Methods.—Most audio-frequency amplifiers employ either resistance or transformer coupling, but other methods can be used and under some conditions have advantages. The most important of these other methods are described below.

Resistance-inductance-coupled Amplifiers.[1]—This type of coupling is illustrated in Fig. 112*a*, and is the same as an ordinary resistance-coupled amplifier except for a small inductance L_c placed in series with the

[1] For further discussion see G. D. Robinson, Theoretical Notes on Certain Features of Television Receiving Circuits, *Proc. I.R.E.*, vol. 21, p. 833, June, 1933.

coupling resistance. An arrangement of this type properly designed will give a more uniform response at high frequencies than is possible with plain resistance coupling. It is also characterized by a rather sharp cut-off at high frequencies, *i.e.*, the amplification falls off relatively rapidly beyond the flat region. These characteristics are illustrated by Fig. 113.

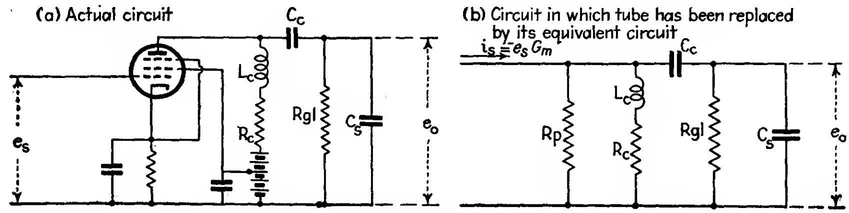


FIG. 112.—Circuit of resistance-inductance-coupled amplifier.

The equivalent circuit of the resistance-inductance-coupled amplifier is shown at Fig. 112*b*. At medium and low frequencies the reactance of the inductance L_c is so small as to be negligible, so that the characteristics are then the same as for the corresponding resistance-coupled case. At high frequencies, however, the inductance tends to resonate with the stray capacities in such a way as to improve the amplification characteristic provided the proper inductance is employed. Analysis shows that, to obtain a flat response up to some particular frequency using pentode

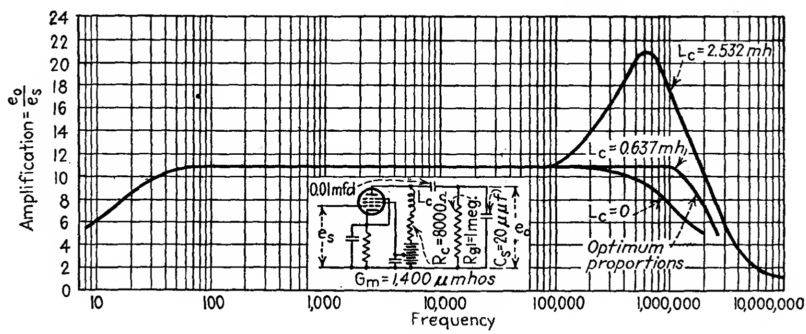


FIG. 113.—Characteristics of resistance-inductance-coupled amplifier, showing how the optimum value of coupling inductance improves the response.

tubes, the coupling resistance should equal the reactance of the shunting capacity at this frequency, and the inductance in series with the coupling resistance should at this same frequency have a reactance equal to half the coupling resistance. The grid-leak resistance should be much higher than the coupling resistance. With these proportions the amplification is substantially constant up to the desired frequency, as shown in Fig. 113, and the phase distortion at high frequencies is also negligible

up to the same point. By using other values of coupling inductance, one can obtain either a rising or a falling characteristic at high frequencies, as illustrated in Fig. 113.

Resistance-coupled Amplifiers with Grid Choke.—In this type of coupling the usual grid-leak resistance of the resistance-coupled amplifier is replaced by a high-impedance choke, as illustrated in Fig. 114. This arrangement has the advantage of reducing to a negligible value the resistance in series with the grid of the tube to which the amplified voltage is delivered, and so permits operation under conditions where the maximum permissible value of grid-leak resistance is low. The grid choke must have a very high inductance, for otherwise it will act as a low-

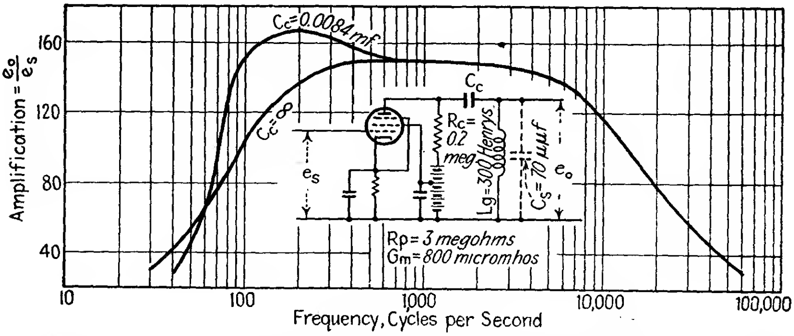


FIG. 114.—Circuit of resistance-coupled amplifier employing grid choke, together with typical amplification characteristics. Note that the character of the low-frequency response can be controlled by the grid choke and coupling-condenser proportions.

impedance shunt at low frequencies. The necessary impedance is relatively easily obtained, however, since the choke carries negligible d-c current and so can be constructed with minimum possible air gap, thereby giving a very large inductance with a small amount of material.

By a proper choice of coupling-condenser capacity and choke inductance in relation to the coupling and plate resistances, it is possible to control the low-frequency response. In particular, the low-frequency response can be made to be substantially constant (but with a slight hump) down to some frequency f_0, and then to fall off rapidly at lower frequencies. This is accomplished by making the inductance L_g of the grid choke such that the choke reactance at the frequency f_0 equals the resistance R formed by plate and coupling resistances in parallel, and then employing a coupling-condenser capacity C_c such as to give series resonance at the frequency f_0 with the inductance L_g. That is,

$$L_g = \frac{R}{2\pi f_0}$$

$$C_c = \frac{1}{(2\pi f_0)^2 L_c} \qquad (99)$$

A larger value of C_c will give a falling off at low frequencies similar to that with ordinary resistance coupling, while a smaller value gives a decidedly peaked low-frequency response.[1] These effects are shown in Fig. 114. Similar characteristics can be obtained by varying the coupling resistance and leaving the remainder of the circuit unchanged. A low coupling resistance tends to give a low-frequency peak, while a high coupling resistance causes a falling off.

Transformer Coupling with Shunt Feed.—Transformer-coupled amplifiers are sometimes arranged as shown in Fig. 115*a*, with a choke[2] and blocking condenser to prevent the d-c plate current from flowing through the transformer primary. This arrangement is known as a shunt-feed

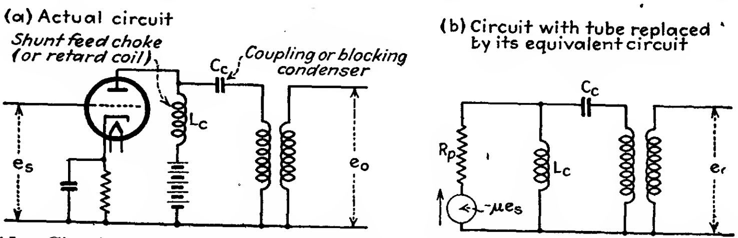


FIG. 115.—Circuit of transformer with shunt feed, together with the equivalent electrical circuit.

circuit and is required when the transformer has a permalloy or other high-permeability core. Shunt-feed arrangements also make it possible to obtain an improved frequency response from transformers using ordinary silicon-steel cores.

The improvement in characteristics obtained with shunt feed results from the fact that the transformer primary has no superimposed direct-current magnetization. This permits the use of a magnetic circuit with a minimum possible air gap, and also makes it possible to use high-permeability core material. Because of the resulting low reluctance of the magnetic circuit, the primary inductance can be made greater in proportion to the physical size of the transformer than it can be when provision must be made for handling direct-current magnetization. At the same time, this improvement in primary inductance is obtained without affecting the leakage inductance or the distributed capacity, and hence without affecting the response at high frequencies. This is because the magnetic leakage paths are largely in air, and so are not

[1] The factor governing the character of the response at low frequencies is the ratio of grid-choke reactance to the resistance formed by plate and coupling resistances in parallel, when the ratio is evaluated at the frequency at which the coupling condenser is in resonance with the grid choke. If the ratio is unity, the conditions of Eq. (99) are realized; if the ratio exceeds unity, the low-frequency response is more peaked; if it is less than unity, the amplification falls off at low frequencies.

[2] A choke used for this purpose is sometimes called a *retard coil*.

affected by the reluctance of the core circuit, while the distributed capacity is likewise independent of core material. Consequently the shunt-feed arrangement gives a smaller leakage coefficient and a widened frequency band.

The equivalent circuit of a shunt-feed transformer arrangement is shown in Fig. 115*b*. If the coupling condenser is sufficiently large to be an effective by-pass at all essential frequencies, it will be noted that the primary inductance of the transformer and the inductance of the shunt-feed choke are essentially in parallel, so that the low-frequency response is the same as though one had an ordinary transformer-coupled circuit in which the primary inductance is the equivalent inductance formed by the actual transformer primary in parallel with the choke. The middle-range and high-frequency responses are the same with shunt feed as with ordinary transformer coupling. It will be observed, however, that the shunting effect of the shunt-feed choke tends to counteract the increased primary inductance of the transformer which results from the elimination of the d-c plate current from the transformer primary. With proper design there is, however, still a very great improvement, since the shunt-feed choke can be physically large, and so can be made to have a very high inductance even though carrying d-c current. At the same time the choke, even when large, contributes nothing to the leakage inductance, whereas a transformer large enough to handle the d-c current and still give a high primary inductance has a high leakage.

The most satisfactory response at low frequencies is obtained by properly proportioning the coupling condenser C_c with respect to the primary inductance L_p of the transformer. A substantially flat low-frequency response down to a frequency f_0 can be obtained by making the primary inductance of the transformer such that the inductive reactance at the frequency f_0 is equal to the plate resistance R_p, making the coupling condenser of such a size that it is in series resonance with the primary inductance of the transformer at the frequency f_0, and using a shunt feed inductance L_c at least one and a half times the transformer primary inductance. These requirements lead to:

$$\begin{aligned} L_p &= \frac{R_p}{2\pi f_0} \\ C_c &= \frac{1}{(2\pi f_0)^2 L_p} \\ L_c &\geqq 1.5 L_p \end{aligned} \tag{100}$$

The resulting frequency response for these proportions is shown in Fig. 116, together with the characteristics obtained with other values of coupling-condenser capacity. It will be noted that the low-frequency response is either peaked or falling, according to whether the reactance

of the transformer at the resonant frequency of C_c and L_p is greater or less than the plate resistance, respectively.

Impedance Coupling.—An impedance-coupled amplifier is a resistance-coupled amplifier in which the coupling resistance has been replaced by a high-impedance choke coil, as shown in Fig. 117*a*. This arrangement has the advantage over resistance coupling that there is negligible direct-current voltage drop in the coupling impedance, so that lower plate voltages and higher plate currents can be used. The cost is much greater, however, and the coupling impedance tends to pick up voltage from stray magnetic fields, so the arrangement is used only in

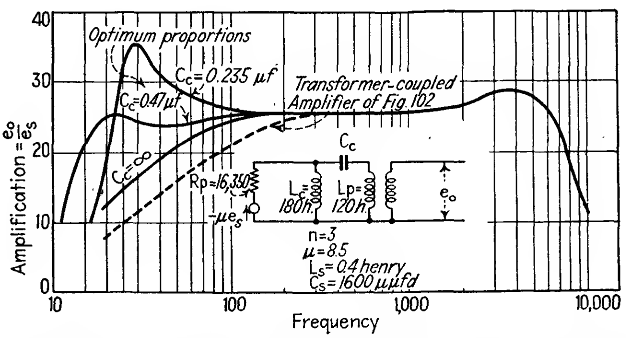


FIG. 116.—Typical amplification characteristics of shunt-feed circuit of Fig. 115 for various coupling-condenser capacities. The transformer has the same characteristics as in Fig. 102 except that the primary inductance is assumed to be tripled as a result of the elimination of direct-current magnetization. For purposes of comparison the curve of Fig. 102 is shown dotted.

special circumstances. The equivalent circuit of an impedance-coupled amplifier is shown in Fig. 117*b*, and the way in which the amplification varies with frequency in a typical triode impedance-coupled amplifier is shown in Fig. 118. In the middle range of frequencies the amplification is high and relatively constant. At low frequencies the amplification falls off because of the drop in the impedance of the choke at low frequencies, while at high frequencies the amplification falls off as a result of the shunting capacities. Equivalent circuits for various frequency ranges are shown in Fig. 117, together with the resulting amplification equations.

Examination of the equations in Fig. 117 shows that, when the plate resistance is very high, as is the case with pentode tubes, the grid-leak resistance must be of the order of magnitude of the coupling resistance that would be used with a resistance-coupled amplifier if a relatively flat frequency response is to be obtained. This is because with very high plate resistance the falling off at low frequencies is determined by the ratio of coupling-inductance reactance to R_1 (=resistance formed by

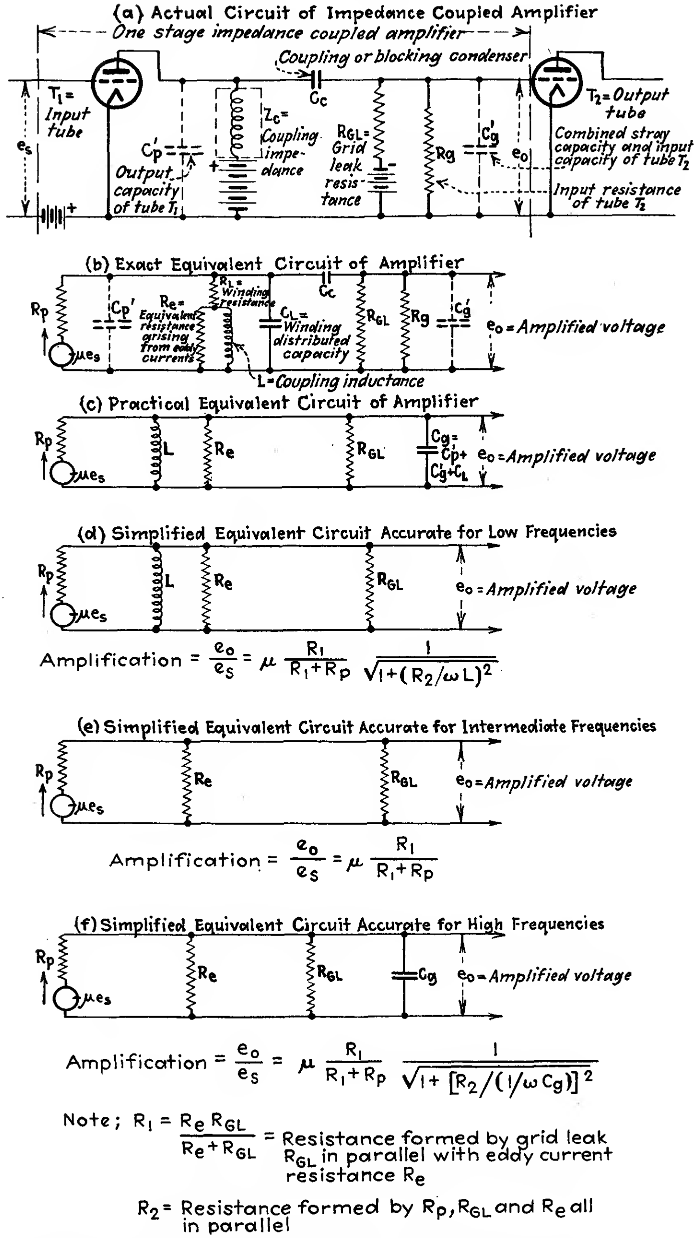


FIG. 117.—Actual circuit of impedance-coupled amplifier together with equivalent electrical circuits.

R_c and R_{gl} in parallel), and at high frequencies by the ratio of shunting-condenser reactance to R_1, with the response being better the lower the grid-leak resistance. This assumes that the coupling condenser is large enough so that the amplification begins to fall off as a result of insufficient coupling impedance some time before the falling off due to excessive condenser reactance becomes appreciable.

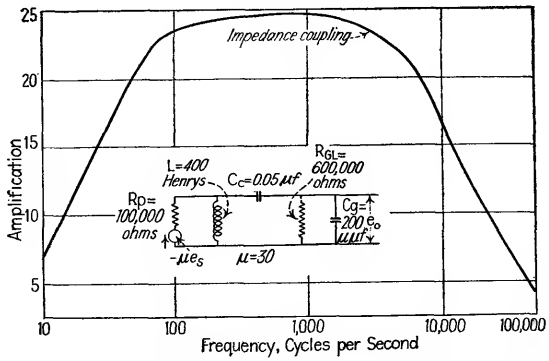


FIG. 118.—Typical amplification curve of an impedance-coupled amplifier.

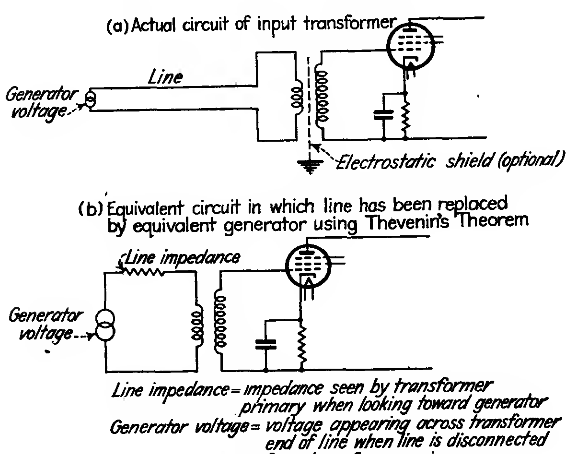


FIG. 119.—Circuit of input transformer.

Input Transformers.—Circumstances commonly arise where it is necessary to couple the grid of a tube to a transmission line. This is usually done with a step-up transformer, which is termed an input transformer. The line that excites the primary of the transformer can be thought of as being equivalent to a generator having a certain internal resistance, as shown in Fig. 119. It will be observed that this is similar in all essential respects to the circuit of an ordinary interstage transformer-

coupled amplifier as discussed above, with the internal resistance of the power source replacing the plate resistance of the tube. Input transformers therefore have characteristics that are similar to interstage transformers, and they are analyzed on the same basis, the only difference in the design being that the primary winding must have a primary inductance in proportion to the line resistance. With low-resistance lines (or sources) the primary inductance and hence the primary turns will be small, and the transformer will have a large step-up ratio; with lines (or sources) having an equivalent resistance of the same order of magnitude as the plate resistance of a tube, the input transformer and the interstage transformer will be interchangeable except for the fact that input transformers normally do not have to be designed to carry d-c current in their primary windings.

The internal impedance of a source of power supplying an input transformer is commonly referred to as the *line impedance*. Thus an input transformer designed to operate with a microphone having an internal impedance of 4 ohms would be said to be designed to operate from a 4-ohm line, since the transmission line that connects the microphone to the transformer would appear to the primary of the transformer to have 4 ohms impedance. Where energy is to be transmitted any distance, line impedances ranging from a few ohms up to 500 or 1000 ohms are commonly used. Higher impedance lines are generally avoided because they are more susceptible to trouble from stray fields.

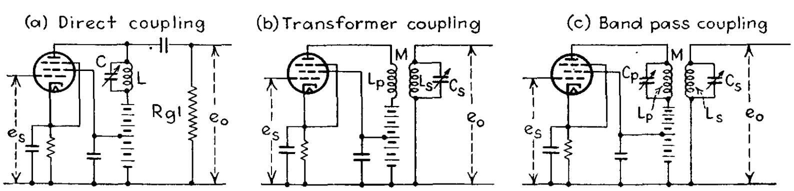


FIG. 120.—Typical tuned-amplifier circuits using pentode tubes.

47. Tuned Voltage Amplifiers.—In a tuned amplifier the load impedance is supplied by a resonant circuit, using parallel resonance to obtain the necessary high impedance. Such amplifiers find their principal use in the amplification of radio frequencies. They are particularly satisfactory for such applications since they can utilize the stray and tube capacities inevitably present to help tune the resonant circuit. They are also selective with respect to frequency, giving an amplification that varies with frequency in much the same way as does the response of an ordinary resonance curve, and thereby making it possible to amplify signals of a desired frequency while eliminating other signals.

Tuned radio-frequency voltage amplifiers practically always employ pentode tubes of the type in which the control grid is very completely

shielded from the plate circuit, as shown in Figs. 59 and 86b. Tubes of this character are commonly called radio-frequency pentodes; they give more gain than triode tubes and also have negligible capacitive coupling between the plate and grid circuits. Screen-grid tubes have many of the advantages of pentode tubes but are not so satisfactory.

A number of typical tuned-amplifier circuit arrangements are shown in Fig. 120. The basic principle of operation in all these is the same, the chief difference between the circuits being the means used to couple the tuned circuit to the plate circuit of the amplifier tube. The analysis of all of these circuits is carried out in the same way, first replacing the tube by its equivalent circuit and then calculating the behavior of the resulting electrical network.

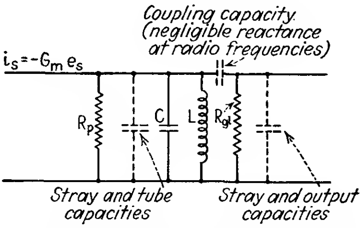


FIG. 121.—Equivalent circuit of tuned amplifier of Fig. 120a, in which the tube has been replaced by an equivalent constant-current generator circuit.

Analysis of Typical Tuned Amplifiers Using Pentode Tubes.—The simplest tuned-amplifier circuit is the direct-coupled arrangement shown in Fig. 120a, and this will accordingly be considered first because it illustrates the principal properties of tuned amplifiers. The exact equivalent circuit is shown in Fig. 121. It will be noted that the stray wiring capacities and the plate-cathode capacity of the amplifier tube, as well as the grid-cathode capacity of the tube to which the output voltage is delivered, help tune the circuit and are therefore not particularly detrimental. From this equivalent circuit the amplification can be written down at once as

$$\text{Amplification} = \frac{e_0}{e_s} = G_m Z_L \quad (101)$$

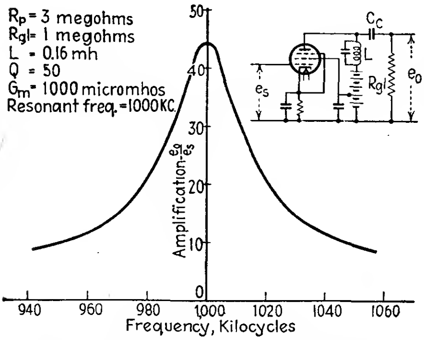


FIG. 122.—Variation of amplification with frequency in a typical tuned amplifier. This curve has the shape of a resonance curve corresponding to a circuit Q slightly less than the actual Q of the resonant circuit.

where G_m is the mutual conductance of the amplifier tube and Z_L is the impedance formed by the tuned circuit, the grid-leak resistance, and the plate resistance, all in parallel.

The results in a typical case are shown in Fig. 122 in the frequency range around resonance. Examination of Fig. 122 and Eq. (101) shows that the amplification curve has the shape of a resonance curve, with the maximum amplification occurring at the frequency at which the tuned

circuit is resonant when taking into account the stray shunting capacities. At this frequency the parallel impedance of the resonant circuit is ωLQ, so that Eq. (101) becomes

$$\left.\begin{array}{l}\text{Amplification at}\\ \text{resonance}\end{array}\right\} = G_m \frac{\omega LQ}{1 + \dfrac{\omega LQ}{R_p} + \dfrac{\omega LQ}{R_{gl}}} \tag{102a}$$

The notation is as illustrated in Fig. 121. With ordinary circuit proportions the grid-leak resistance and plate resistance are both very much higher than the parallel impedance of the resonant circuit, so that to a good approximation one can rewrite Eq. (102a) as

$$\left.\begin{array}{l}\text{Approximate amplification}\\ \text{at resonance}\end{array}\right\} = G_m \omega LQ \tag{102b}$$

The curve of amplification as a function of frequency will be found to have a shape corresponding to an effective Q lower than the actual Q of the resonant circuit. The ratio of this effective Q of the amplification curve to the actual Q of the tuned circuit depends upon the grid-leak and plate resistances associated with the tuned circuit; it is given by the equation[1]

$$\frac{\text{Effective } Q \text{ of amplification curve}}{\text{Actual } Q \text{ of tuned circuit}} = \frac{1}{1 + \dfrac{\omega LQ}{R_p} + \dfrac{\omega LQ}{R_{gl}}} \tag{103}$$

The transformer-coupled circuit of Fig. 120b behaves in much the same way as does the direct-coupled circuit of Fig. 120a, although differing in details of analysis. The exact equivalent circuit of the transformer-coupled arrangement is shown in Fig. 123a for both the constant-voltage and constant-current forms of the equivalent tube circuit. These equivalent circuits are relatively complicated as a result of the various stray and tube capacities. However, if the primary inductance is not too large and if the coefficient of coupling is reasonably high, then the principal effect of these capacities is to assist in the tuning of the resonant circuit. This gives the equivalent circuits of Fig. 123b. The constant-voltage generator form of this circuit has already been discussed in Sec. 17 as an example of a typical coupled circuit, and it was found that the response in the secondary circuit had the shape of a resonance curve corresponding to an effective Q less than the actual Q of the tuned circuit. The amplification of the transformer-coupled circuit at resonance is readily shown to be

[1] This ratio is simply the ratio of the resonant impedance of the tuned circuit to the impedance formed by the combination of tuned circuit, grid-leak resistance, and plate resistance, all in parallel.

$$\left.\begin{array}{l}\text{Amplification at}\\ \text{resonance}\end{array}\right\} = G_m \frac{\omega MQ}{1 + \frac{(\omega M)^2/R_s}{R_p}} \tag{104a}$$

In pentode tubes R_p is extremely high, so that to a good approximation

$$\left.\begin{array}{l}\text{Approximate amplification}\\ \text{at resonance}\end{array}\right\} = G_m\omega MQ \tag{104b}$$

The ratio of the effective Q of the amplification curve to the actual Q of the tuned circuit is given by[1]

$$\frac{\text{Effective } Q \text{ of amplification curve}}{\text{Actual } Q \text{ of tuned circuit}} = \frac{1}{1 + \frac{(\omega M)^2/R_s}{R_p}} \tag{105}$$

When the plate resistance is very large, as in the case of pentode amplifier

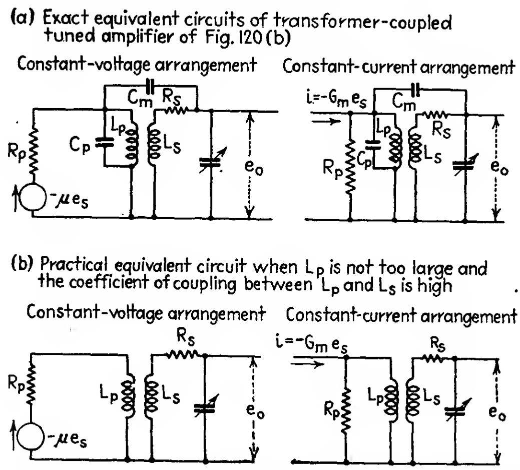


FIG. 123.—Exact and practical equivalent circuits of transformer-coupled amplifier of Fig. 120b.

tubes, the effective Q of the amplification curve is substantially the same as the Q of the resonant circuit.

An examination of Eq. (104a) shows that the mutual inductance controls the equivalent impedance which is coupled into the plate circuit of the tube by the tuned circuit and hence controls the amplification. In the case of tuned amplifiers employing triode tubes, the transformer

[1] This equation results from the fact that the response curve has the shape of a resonance curve corresponding to a secondary resistance of $R_s + \frac{(\omega M)^2}{R_p}$ instead of the value R_s actually present. The right-hand side of Eq. (105) is simply the inverse ratio of these two resistances.

coupling provides a means of matching the load impedance provided by the tuned circuit to the plate resistance of the tube in such a way as to obtain the maximum possible amplification, as discussed below. With pentode tubes the problem of matching does not occur because the plate resistance of the tube is so much more than any load resistance that can conceivably be coupled into the plate circuit that the amplification will be greater the larger the coupling.

Tuned Amplifiers with Complex Coupling.[1]—When the resonant frequency of the tuned circuit is adjusted by varying the capacity, the

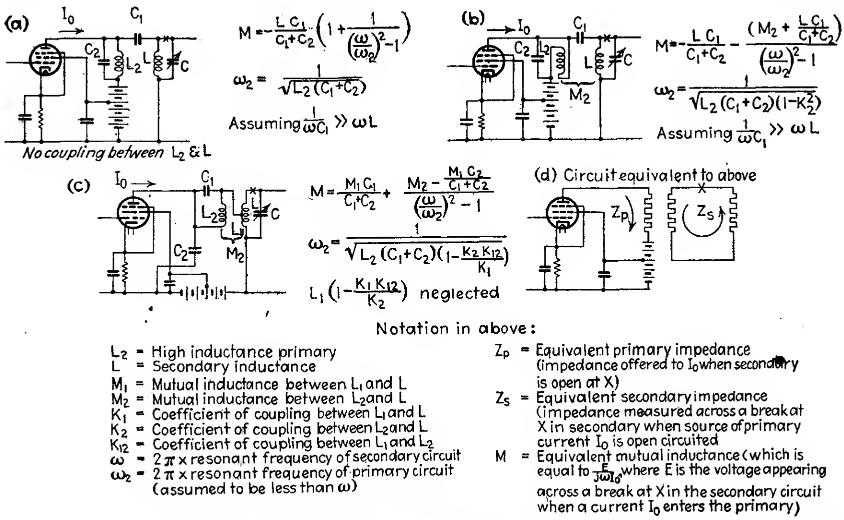


FIG. 124.—Typical tuned amplifiers with complex coupling, together with equations for the equivalent mutual inductance M and the equivalent electrical circuit representing the complex network.

direct- and transformer-coupled arrangements of Fig. 120 give an amplification that is roughly proportional to the resonant frequency, as is evident when the equations of amplification at resonance are examined and it is remembered that Q is roughly independent of frequency. This behavior is undesirable under ordinary circumstances, and it can be overcome by coupling the amplifier tube to the tuned circuit by a complex network such that the way in which the amplification varies with the frequency can be controlled by the circuit constants.

Typical examples of complex coupling arrangements that can be used to accomplish this result are given in Fig. 124. These circuits are relatively complicated, but, as has already been explained in Sec. 18, they can in every case be simplified to the equivalent transformer-coupled

[1] A very complete discussion of this subject is given by Harold A. Wheeler and W. A. MacDonald, Theory and Operation of Tuned Radio-frequency Coupling Systems, *Proc. I.R.E.*, vol. 19, p. 738, May, 1931.

arrangement shown in Figs. 40g and 124d. In this equivalent circuit one postulates a primary impedance Z_p, which is the impedance actually observed across the primary terminals when looking into the network with the secondary circuit open. Similarly the secondary is assumed to have the impedance Z_s, which is the impedance that is observed when the secondary circuit is opened at the point X and the primary terminals are left open-circuited. The mutual inductance M, which is assumed to exist between the primary and secondary impedances, is determined from the fact that, when a current I_0 flows into the primary terminals, the voltage that appears across a break at X in the secondary circuit is $-j\omega MI_0$. In speaking of the quantity M in this case as an "equivalent mutual inductance," it must be realized that what is really being referred to is the equivalent coupling between the circuits. When this coupling is capacitive, the "mutual inductance" will actually represent a mutual capacity, which is indicated mathematically by M having a negative sign. When there is combined reactive and resistive coupling, M will be a complex quantity.

The method of procedure in analyzing complex coupling can be understood by considering a simple case such as that shown in Fig. 124a. As this circuit is commonly encountered, the primary L_2 has a high inductance compared with the secondary inductance L. The capacity C_2, representing the distributed capacity of L_2 plus the plate-cathode capacity of the amplifier tube, is made large enough to be in resonance with the inductance L_2 at a frequency slightly below the lowest frequency to be amplified. The primary inductance L_2 normally has no coupling at all to the secondary inductance L, the entire coupling being provided by the capacity C_1, which is very small, often only 2 or 3 $\mu\mu$fds. provided by stray capacity.

In an arrangement of this sort one is primarily interested in the effective mutual inductance M and the way it varies with frequency. The only other action the complex coupling arrangement can have is to influence slightly the tuning of the secondary circuit. The first step in calculating M is to determine the voltage produced across a break X in the secondary when a current I_0 enters the primary. Assuming $1/\omega C_1 >> \omega L$, this is[1]

[1] This is derived as follows: If $1/\omega C_1 >> \omega L$, then the impedance offered to I_0 is essentially that of the parallel impedance of a circuit having an inductance L and a capacity $(C_1 + C_2)$; thus it is

$$\left.\begin{array}{l}\text{Impedance}\\ \text{offered by}\\ \text{primary to } I_0\end{array}\right\} = Z_p = \frac{j\omega L_2/j\omega(C_1 + C_2)}{j\left(\omega L_2 - \dfrac{1}{\omega(C_1 + C_2)}\right)} = \frac{L_2}{(C_1 + C_2)j\omega L_2\left[1 - \dfrac{1}{\omega^2 L_2(C_1 + C_2)}\right]}$$

If one denotes by $\omega_2/2\pi$ the resonant frequency of the primary (*i.e.*, $\omega_2 L_2 =$

$$\left.\begin{matrix}\text{Open circuit secondary}\\ \text{voltage}\end{matrix}\right\} = j\omega I_0 \frac{LC_1}{C_1 + C_2}\left[1 + \frac{1}{(\omega/\omega_2)^2 - 1}\right] \quad (106)$$

Dividing this by $-j\omega I_0$ gives the value for M shown in Fig. 124a. The result is an effective mutual inductance, and hence an effective coupling coefficient, which varies with frequency in a manner determined by the resonant frequency $\omega_2/2\pi$ of the primary inductance. The variation in mutual inductance M with frequency in an actual case for several values of ω_2 is given in Fig. 125, and it is apparent that by suitable choice of circuit constants it is possible to counteract in a large measure the tendency for the amplification to vary with the resonant frequency of the secondary.

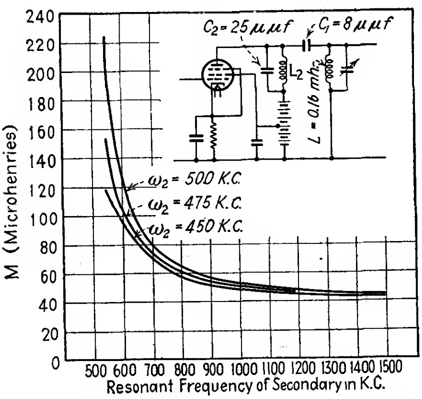


FIG. 125.—Variation of equivalent mutual inductance M with frequency for several circuit proportions of the complex coupling system of Fig. 124a. For constant amplification with a fixed value of Q, the value of M should vary inversely with frequency, and it is apparent that this ideal can be roughly approximated by the use of complex coupling.

The other circuit arrangements of Fig. 124 are somewhat more difficult to analyze than the example just considered, but they lead in the end to similar results. The value of the equivalent coupling M in each case is given on the figure, and the detailed analysis is to be found in the literature.[1]

Band-pass Amplifiers.—A band-pass characteristic such as shown in Fig. 41b can be realized in an amplifier by using in place of a single reso-

---

$1/\omega_2(C_1 + C_2)$), then $\omega^2 L_2(C_1 + C_2) = (\omega/\omega_2)^2$ and

$$\left.\begin{matrix}\text{Impedance offered}\\ \text{by primary to } I_0\end{matrix}\right\} = Z_p = \frac{1}{j\omega(C_1 + C_2)\left[1 - \frac{1}{(\omega/\omega_2)^2}\right]}$$

The voltage developed across the primary is the product Z_pI_0 of this impedance and I_0; if $1/\omega C_1 >> \omega L$, the voltage across L (which is also the voltage appearing across a break at X in the secondary) $= Z_pI_0\frac{j\omega L}{1/j\omega C_1} = -Z_pI_0\omega^2 LC_1 =$

$$(j\omega)^2 I_0 LC_1 \frac{1}{j\omega(C_1 + C_2)\left[1 - \frac{1}{(\omega/\omega_2)^2}\right]} = j\omega I_0 \frac{LC_1}{(C_1 + C_2)\left[1 - \frac{1}{(\omega/\omega_2)^2}\right]} =$$

$j\omega I_0 \frac{LC_1(\omega/\omega_2)^2}{(C_1 + C_2)[(\omega/\omega_2)^2 - 1]}$. This last form is Eq. (106) with a rearrangement of terms.

[1] *Ibid.*

nant circuit two resonant circuits tuned to the same frequency and suitably coupled as shown in Fig. 126*a*. This arrangement is commonly used in intermediate-frequency amplifiers and is very desirable for the amplification of modulated waves because it can provide substantially constant amplification for all the essential side-band frequencies contained in the wave, while at the same time discriminating sharply against frequencies that are outside the pass band. In contrast with this, a tuned amplifier composed of a single resonant circuit has a rounded-off response curve, and so cannot discriminate against interfering frequencies just outside the desired frequency band without at the same time discriminating against the higher frequency side-band components.

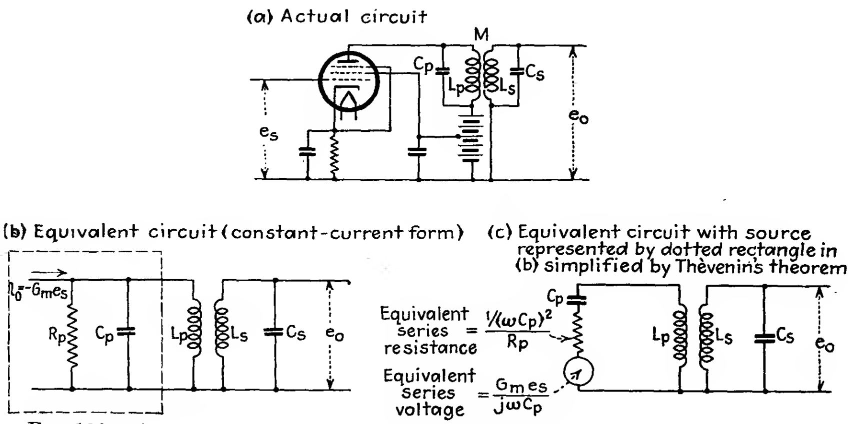


FIG. 126.—Actual circuit of band-pass amplifier, together with equivalent circuit.

The equivalent circuit of a band-pass amplifier is shown in Fig. 126*b*. Although this circuit appears complicated, it can be analyzed in a very simple manner by using Thévenin's theorem to replace the actual source of voltage by an equivalent voltage acting in series with the primary circuit. Thus, applying Thévenin's theorem to the network enclosed in the dotted rectangle in Fig. 126*b*, the result is the equivalent circuit shown at Fig. 126*c*, which is identical with the band-pass circuits discussed in Secs. 17 and 19. Because the plate resistance R_p of the tube is always much greater than the reactance of the condenser C_p even when triode tubes are employed, the equivalent voltage E is given by the formula

$$\left.\begin{array}{l}\text{Equivalent voltage}\\ \text{that can be considered}\\ \text{as acting in series}\\ \text{with the primary}\end{array}\right\} = e_s G_m \left(\frac{1}{j\omega C_p}\right) \qquad (107)$$

Assuming $R_p >> 1/\omega C_p$, the equivalent internal impedance of the generator in the dotted rectangle is substantially equivalent to the capacity C_p in series with a small added resistance of $(1/\omega C_p)^2/R_p$ ohms.[1] It is apparent from these considerations that the only effect which is produced upon the band-pass characteristic by associating a pair of tuned circuits with the tube, particularly a pentode, is to reduce slightly the effective primary Q. The band width, shape of response curve, etc., are hardly modified at all.

The amplification can be calculated readily with the aid of Eqs. (45) and (48*a*) by taking advantage of the fact that the equivalent voltage acting in series with the primary circuit is given by Eq. (107), while the output voltage is equal to the secondary current times the reactance of the secondary condenser. At the common resonant frequency this gives

$$\left.\begin{array}{l}\text{Amplification at}\\ \text{resonant frequency}\end{array}\right\} = G_m k \frac{\omega_0\sqrt{L_s L_p}}{k^2 + \dfrac{1}{Q_p Q_s}} \tag{108a}$$

where

G_m = mutual conductance of tube

k = coefficient of coupling between primary and secondary inductances

ω_0 = 2π times resonant frequency

Q_p = $\omega L_p/R_p$ for primary circuit, taking into account any equivalent resistance that may be added by the plate resistance of the tube.

Q_s = $\omega L_s/R_s$ for secondary circuit

L_p, L_s = primary and secondary inductances, respectively.

The amplification is maximum when the coefficient of coupling has the critical value as defined by Eq. (47), *i.e.*, $k = 1/\sqrt{Q_p Q_s}$, and for this condition becomes

$$\left.\begin{array}{l}\text{Maximum possible}\\ \text{amplification}\end{array}\right\} = G_m \frac{\omega_0\sqrt{L_p L_s}\sqrt{Q_p Q_s}}{2} \tag{108b}$$

Comparison of Eqs. (108*b*) and (102*a*) shows that, when the primary and secondary circuits are identical and the plate and grid-leak resistances are very high, the gain obtained with the band-pass arrangement is exactly half that which is obtained when a single tuned circuit is employed. When the coefficient of coupling is greater than the critical value, the amplification curve becomes double humped, exactly as do the curves of secondary current in the band-pass filters discussed in Sec. 19. The

[1] This results from noting that by Thévenin's theorem the internal impedance of the equivalent generator consists of a capacity C_p shunted by a resistance R_p. This shunt resistance is then transformed into an equivalent series resistance by Eq. (19).

amplification at these peaks is substantially equal to the amplification as given by Eq. (108*b*), and the separation of the peaks is determined by the factors discussed in connection with Eqs. (49*a*) and (49*b*). The amplification at the resonant frequency is still given by Eq. (108*a*), irrespective of whether or not double humps occur.

Tuned Amplifiers Employing Triode Tubes.—The first tuned amplifiers employed triode tubes, but triodes are now completely displaced by radio-frequency pentodes. This is because pentodes give more amplification, make it possible to obtain better selectivity, and have practically no direct electrostatic capacity between grid and plate electrodes to transfer energy between input and output circuits. As a consequence, triode tubes are of little importance as tuned voltage amplifiers, although they find a very important use as tuned *power* amplifiers, as discussed in the next chapter.

The analysis of tuned amplifiers employing triode tubes is carried out in exactly the same way as with pentode tubes, and the same equivalent circuits apply. The constant-voltage generator form of these circuits is usually preferable with triodes, however, whereas the constant-current generator form is simplest when dealing with pentodes. The only essential difference that occurs with triode amplifiers arises from the fact that the plate resistance of a triode is relatively low and so becomes an important factor. In particular, the low plate resistance makes the effective Q of the amplification curve appreciably less than the actual Q of the resonant circuit, as is apparent from Eq. (105). There is also an optimum coupling between the tuned circuit and plate of the tube for maximum amplification. This optimum condition can be shown by Eq. (105) to be when the resistance that the tuned secondary couples into the plate circuit at resonance equals the plate resistance of the tube. Under such conditions the effective Q of the amplification curve is exactly one-half the actual Q of the tuned circuit.

Design of Tuned Amplifiers.—The design of ordinary tuned amplifiers can be carried out in a straightforward manner. The tubes normally used are radio-frequency pentodes, generally of the variable-mu type, and are well standardized. Tubes are operated at the highest mutual conductance that it is practicable to realize at the potentials available and still keep within the tube rating. The screen electrode should be by-passed to the cathode or to ground by a suitable condenser. The tuning-coil inductance is usually determined by such considerations as physical size and broadness of response necessary to prevent undue discrimination against the higher side-band frequencies, while at the same time keeping in mind that the gain will be increased by a high Q and a high ratio of inductance to capacity. Any convenient coupling method will be satisfactory provided it is designed with the consideration that, although

the gain per stage is increased by coupling the tuned circuit very closely to the plate circuit, a high gain per stage increases the likelihood of trouble from oscillations, as discussed in Sec. 51. When the resonant frequency of the tuned circuit is to be varied, a complex coupling arrangement designed to give substantially constant amplification over a range of frequencies is desirable.

Selectivity and Selectivity Curves of Tuned Amplifiers.—The ability of a tuned amplifier to discriminate against interfering signals is often fully as important as the gain of the amplifier at resonance. When the load impedance contains only a single tuned circuit, the discrimination

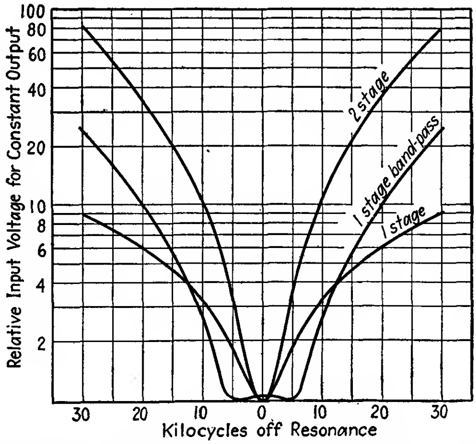


FIG. 127.—Selectivity curves of tuned amplifiers, in which the ordinates show relative input voltages required at different frequencies to maintain the output voltage constant.

against frequencies appreciably off resonance is given approximately by the equation

$$\frac{\text{Amplification when far off resonance}}{\text{Amplification at resonance}} = \frac{1}{Q(1 - \gamma^2)} \qquad (109a)$$

where Q is the effective Q of the amplification curve and γ equals the ratio of the actual frequency to the resonant frequency. When the load impedance is a band-pass filter such as is shown in Fig. 126, the discrimination against frequencies far off resonance, for couplings equal to or greater than the critical value, is given approximately by

$$\frac{\text{Amplification when far off resonance}}{\text{Maximum amplification at coupling peak}} = \frac{2k^2}{\gamma(1 - \gamma^2)^2} \qquad (109b)$$

where k is the coefficient of coupling between primary and secondary circuits and γ is the ratio of actual frequency to the resonant frequency. Equation (109*b*) is derived from Eqs. (45), (47), (48*b*), and (107) by neglecting resistance terms when far off resonance.

The effectiveness of an amplifier in discriminating against frequencies off resonance can be most satisfactorily shown by plotting a curve of relative amplification in the manner shown in Fig. 127, which gives the relative input required to maintain the output constant as the frequency of the input is varied. The input required to produce the output at the frequency of maximum amplification is used as the reference value. Thus, in the band-pass case in Fig. 127, it takes an input signal 25 times as strong to produce a given output when the carrier is 30 kc off resonance as it does when the carrier is exactly at resonance. Included in Fig. 127 are the selectivity curves of one and two stage amplifiers employing single tuned circuits, and also a single stage of band-pass amplification. The superiority of the band-pass amplifier in its ability to discriminate against frequencies appreciably off resonance, while at the same time responding uniformly to a band of frequencies about resonance, is clearly evident.

Problems

1. *a.* An alternating potential of 2 volts having a frequency of 1000 cycles is applied to the grid of a triode having $\mu = 14$, $R_p = 10{,}000$, and a resistance load of 12,000 ohms. Calculate (1) alternating current in the load, (2) alternating voltage across the load, (3) alternating power dissipated in the load, and (4) ratio of voltage across load to the grid voltage.

b. Repeat the above for a load consisting of a 2-henry inductance.

c. Repeat (*a*) and (*b*) using the constant-current form of the equivalent circuit.

2. *a.* Design a resistance-coupled amplifier using the pentode tube of Figs. 62 to 65, when a plate-supply potential of 350 volts is available and when only the audio-frequency range is to be covered. In this design determine approximate screen voltage and plate current at the actual operating point, bias resistance, and bias by-pass condenser as well as the coupling resistance, grid-leak resistance, and coupling condenser.

b. Calculate and plot the amplification as a function of frequency by making an estimate of the mutual conductance and plate resistance at the operating point and assuming the total shunting capacity to be 35 $\mu\mu$f.

3. Repeat Prob. 2, but design the amplifier for television purposes requiring a response that does not fall to less than 70 per cent of the maximum for a frequency range of 20 to 400,000 cycles. Assume that the shunting capacities are reduced to 15 $\mu\mu$f in this case, and use plain resistance coupling without the addition of an inductance.

4. When the plate-load resistance or the screen voltage is so high that a virtual cathode forms in front of the suppressor grid in a pentode resistance-coupled amplifier, what will be the effect on the amplification obtained?

5. *a.* Design a resistance-coupled amplifier to cover the speech range, using a Type 75 (or 2A6) high-mu triode when a plate-supply voltage of 300 volts is available. Include in this design the determination of the actual operating point, using characteristic curves published in a tube manual.

b. Calculate the expected amplification as a function of frequency, using data on tube constants obtained from a tube manual and assuming that the shunting capacities are 100 $\mu\mu$f.

6. Check the curve of Fig. 102 by recalculating with the data given.

7. A transformer-coupled amplifier is to give a mid-frequency gain of 40 when operated with a 56 tube, and it must cover the frequency range of 50 to 14,000 cycles with an amplification that does not vary by over 20 per cent from 40. Specify the proper primary and leakage inductances, the turn ratio, and the equivalent distributed capacity.

8. A certain transformer is found to have the following characteristics upon test:

Primary inductance at rated primary current	=	30 h
Primary inductance with secondary shorted	=	0.3 h
Ratio of transformation (voltage ratio at low frequencies)	=	4.0
Voltage ratio at 4000 cycles (no output tube)	=	6.0
R_e so large as to be of no consequence		
Primary direct-current resistance	=	500 ohms
R_s (actual value before reducing to unity-turn ratio)	=	10,000 ohms

This transformer is to be used with a 56 tube having an amplification factor of 14 and a plate resistance of 9500 ohms. The input capacity of the following tube is estimated as 75 $\mu\mu$f. Calculate and plot the way in which the amplification would be expected to vary with frequency, and comment as to how the characteristic might be improved.

9. Calculate the amplification curve that would be obtained in Fig. 102 when a 15,000-ohm resistance is shunted across the transformer primary.

10. Derive Eqs. (95) and (96).

11. Redesign the amplifier of Prob. 3 as a resistance-inductance-coupled amplifier, and for comparison replot the resistance-coupled curve of amplification along with the curve for resistance-inductance coupling.

12. A transformer is to couple a 500-ohm line to the grid of a tube. Assuming that the turn ratio is to be 15, specify the minimum allowable primary inductance, the maximum allowable leakage inductance (referred to the primary), and the desired series-resonance frequency, if the response is to be within 30 per cent of the mid-frequency response over the frequency range of 80 to 7000 cycles.

13. Check the curve of Fig. 118 by recalculating from the circuit data given on the figure.

14. Check the amplification curve of Fig. 122 by recalculating.

15. *a.* Design a tuned amplifier using the circuit of Fig. 120*a*, a 6D6 tube, and a coil for the tuned circuit which is the coil of No. 28 wire in Fig. 19.

b. Calculate the curve of amplification as a function of frequency up to 30 kc on either side of resonance when the coil is tuned to 650 and 1400 kc.

c. Repeat (*a*) and (*b*) for the circuit of Fig. 120*b* when the mutual inductance between primary and secondary is 100 μh.

16. *a.* Calculate the amplification at resonance as a function of resonant frequency over the range of 550 to 1500 kc, for Prob. 15*a*.

b. Design a complex coupling system following one of the arrangements of Fig. 124, so that the amplification will be approximately the same at 700 and 1250 kc, and plot the resulting amplification as a function of frequency on the same axes as the curve of (*a*).

17. In a typical intermediate-frequency amplifier of the band-pass type shown in Fig. 126, the primary and secondary coils each have inductances of 4 mh and are tuned to 260 kc. The coil Q's are 50 and the coupling is adjusted so that $k = 0.03$. If a 6D6 amplifier tube is used, calculate and plot the curve of amplification as a function of frequency.

18. Repeat Prob. 17 for $k = 0.015$ and for $k = 0.05$, and plot on the same curve sheet as used in Prob. 16. Also plot the same three curves in the form of selectivity curves of the type shown in Fig. 127.

CHAPTER VI

VACUUM-TUBE AMPLIFIERS (*Continued*)

48. Methods of Volume Control.—It is nearly always necessary to provide means of controlling the gain of an amplifier so that the output can be kept at a desired level irrespective of the signal voltage being amplified. Thus in the case of a radio receiver it is desired to control the loudness of the loud-speaker output without regard to the strength of the radio signal. The volume-control arrangement for accomplishing this should be such that the gain setting has little or no effect upon the frequency response[1] and does not introduce amplitude distortion.

Volume Control in Audio-frequency Amplifiers.—The standard method of controlling volume in resistance-coupled amplifiers is shown in Fig. 128*a*, where the grid leak is supplied by a high-resistance potentiometer. The only effect produced upon the amplifier characteristics by such a volume control is a slight improvement in the high-frequency response at low volume settings.

The control of volume in transformer-coupled amplifiers is complicated by the fact that the circuit involving the transformer must be very carefully proportioned to give proper frequency response. The best arrangement for controlling volume when transformers are used is shown in Fig. 128*b*, where a high-resistance potentiometer is connected across the transformer secondary. This arrangement is not entirely satisfactory, however, because it introduces a high resistance in the grid circuit at certain volume-control settings and because it modifies the characteristics of the amplifier. In particular, there is reduction in the amplification and alteration of the frequency-response characteristic. When an amplifier contains both resistance- and transformer-coupled stages, it is common practice to control the volume in one of the resistance-coupled stages.

Volume Control in Tuned Amplifiers.—The volume of tuned amplifiers using pentode tubes is practically always controlled by varying the mutual conductance of the amplifier tube, either by changing the control-

[1] The only exception to this is under conditions where it is desired to compensate for the fact that the sensitivity of the ear to high and low pitches relative to the middle range of frequencies is less for weak sounds then for loud sounds (see Sec. 148). This means that for proper reproduction at low output levels, frequency distortion tending to favor the high and low frequencies is desirable. Volume controls which provide this sort of characteristic are said to be "tone compensated."

grid bias or the screen-grid potential. This method of controlling volume does not affect the shape of the frequency-response characteristic, and it has the advantage of permitting the amplification of several stages to be controlled simultaneously. When the gain of a tuned amplifier is to be controlled, variable-mu tubes are practically always employed since these eliminate amplitude distortion and cross-talk troubles at low volume settings (see Sec. 55).

Automatic Volume Control.—Circumstances sometimes arise where it is desired to maintain the output automatically constant irrespective of the input to the amplifier. Thus in a radio receiver it is desirable to

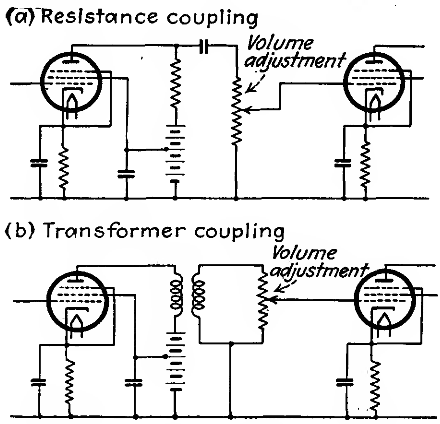


FIG. 128.—Volume-control arrangements normally used in resistance- and transformer-coupled amplifiers.

maintain constant output when the signal being received fades through wide ranges in amplitude.

Automatic volume control of this type is accomplished by rectifying a portion of the output voltage, passing the rectified d-c current through a resistance, and utilizing the voltage thereby developed across the resistance to control the volume in such a way as to tend to maintain the output constant. With tuned amplifiers the customary practice is to employ variable-mu tubes, with the control grids of the various amplifier stages biased negatively by the rectified current. When the signal becomes larger, the output voltage then tends to increase, but this increases the negative bias that the rectified output produces and thereby reduces the gain of the amplifier, tending to counteract the effect of the increased input signal. Detailed discussion of automatic volume-control arrangements used in radio receivers is given in Sec. 109.

Automatic volume control is used in audio-frequency amplifiers only under special circumstances, such as in laboratory oscillators, etc. The simplest method of applying automatic volume control to

audio-frequency amplifiers is to utilize the direct-current voltage developed by the rectified current to control the bias on the grid of a resistance-coupled variable-mu pentode stage operating at low power levels.

Volume Expansion and Contraction.—Under many circumstances it is desirable or necessary to restrict the volume range in audio-frequency circuits. Thus, in recording sound on phonograph records, the loudest passages must be reduced in intensity to prevent the needle from cutting into adjacent grooves, while the weaker passages must have their intensity increased in order that they will not be lost in the background of hiss. A similar situation exists in connection with broadcast transmitters, where it is necessary to prevent, on the one hand, overloading of the transmitter on high-intensity peaks and, on the other hand, the loss of the sound in the background noise and hum during weak passages.

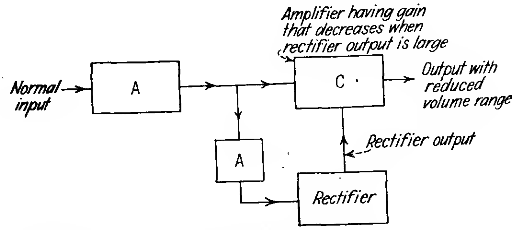


FIG. 129.—Schematic diagram of a volume compressor.

This situation has led to the development of devices for automatically restricting or expanding the volume range. Such volume expanders and contractors are essentially amplifiers having automatic volume control arranged in such a way that the average volume level over a short period of time is used to control the gain of the amplifier. A typical schematic arrangement of a volume compressor is shown in Fig. 129. Here the signal is amplified by AA in the normal manner. A portion of the output of this amplifier is rectified, and a direct-current output voltage is obtained that is proportional to the average amplitude of the output over a small fraction of a second. This direct-current voltage is then used to control the gain of the amplifier C in such a way that the amplification increases when the rectified output is small and decreases when the rectified output is large. A volume expander operates upon the same principle except that the output of the rectifier is used to control the amplifier C in such a way as to increase the amplification during the loud passages and to decrease it during the weak passages. The principal problem involved in volume expanders and compressors is to arrange the rectifier circuit so that the expansion or contraction can rise and fall at the proper rate to give the most satisfactory result. Detailed circuit

arrangements for meeting different requirements are to be found in the literature.[1]

49. Noise, Hum, and Microphonic Action.—All amplifiers give some output even when there is no input voltage. Such output is referred to as hum, noise, or microphonic action, depending upon its origin.

Hum.—Hum is the result of induction from neighboring circuits carrying 60-cycle alternating currents; it is particularly troublesome in audio-frequency amplifiers having high amplification because any hum that is induced in the first stages of such an amplifier will be greatly amplified by subsequent stages. Hum is less troublesome in radio-frequency amplifiers, but cannot be neglected because, even though the induced hum voltages are not amplified, they superimpose voltages upon the tubes that can vary the amplification of radio-frequency signals. In this way the hum may become modulated upon the high frequency.

The possible sources of hum are stray magnetic fields, stray electrostatic fields, alternating current in the heater or filament of the tube, and improperly filtered power-supply systems. Proper design of a power-supply system will readily eliminate all hum from this cause. The problem of avoiding hum while using alternating current to heat the cathode of the tube is discussed in Sec. 92 and can be handled by a suitable choice of tubes. The hum that is the most difficult to handle arises from stray magnetic and electrostatic fields.

Magnetic fields induce voltages in coils, particularly interstage and input audio-frequency transformers, induce voltages in loops that may exist if the wiring is improperly arranged, and may even affect the flow of electrons between the cathode and plate of the tube. The principal sources of magnetic fields in the vicinity of an amplifier are the power transformer, the leads carrying the filament or heater current of the tubes, and the filter chokes. Trouble from the filament leads can usually be eliminated by twisting these, and in some cases by changing their location. Stray fields from power transformers and filter chokes can be reduced by proper orientation, by making the spacing between these pieces of equipment and the first stages of the amplifier as great as possible, and by designing the transformers and chokes to have low leakage flux (achieved by using low flux densities and the smallest possible air gap in the core).

Electrostatic fields cause trouble with parts of the amplifier having a high impedance to ground, since any electrostatically induced current in

[1] C. M. Sinnett, Practical Volume Expansion, *Electronics*, vol. 8, p. 14, November, 1935.

R. C. Mathes and S. B. Wright, The "Compandor"—An Aid against Radio Static, *Elec. Eng.*, vol. 53, p. 860, June, 1934.

flowing to ground will produce a hum voltage that is proportional to the impedance between this part of the circuit and the ground. In particular, when the grid of an audio-frequency amplifier tube is left disconnected, or is grounded through a very high resistance, the impedance between grid and ground is so high that an especially large hum voltage will be developed between grid and ground by the electrostatic fields of near-by lighting circuits. To eliminate trouble from electrostatic pick-up, the tubes should be shielded, and the grid leads should be either shielded or made very short. In audio-frequency amplifiers having very high over-all gain, it is usually necessary to enclose the first one or two stages in a metal box, not leaving even a fraction of an inch of the input circuit (including wires, plugs, etc.) without electrostatic shielding. Filament wires carrying alternating current to the first stages of a high-gain amplifier should be enclosed in grounded metal braid. It is also essential that the chassis of high-gain amplifiers be grounded to water pipes or to a stake driven into moist earth unless all power, filament, input, and output transformers have electrostatic shields.

The most difficult hum problem encountered in audio-frequency amplifiers is from voltages induced in transformers associated with the first few stages of amplification, particularly the input stage. Here the signal being amplified is very small and it takes very little induced hum voltage to be comparable to the signal. The best method of handling this problem is to use resistance coupling throughout, and, if possible, to dispense with the input transformer to the first stage of amplification by using direct coupling, even if this results in considerable loss in amplification.

If an input transformer must be used, it should then be spaced as far as possible from the power transformer, should be in a position that experiment indicates will cause the least hum pick-up, and should be magnetically shielded. Enclosing the transformer in a cast-iron case provides some magnetic shielding, while heat-treated permalloy shields are still more effective. Heavy copper shields have been used with success in some instances. When all other means of eliminating hum from the input transformer fail, it is always possible to place the transformer and the first stage or two of amplification in a separate unit, which can be located some distance away from the power transformer and energized through cables. The line connecting the microphone or other generator to the input transformer should be preferably of low impedance, *i.e.*, 600 ohms or less, and should consist of a twisted pair of wires electrostatically shielded. It is also desirable, although not essential, that the input transformer be provided with an electrostatic shield between primary and secondary.

Microphonic Noises.[1]—When a tube is jarred, the electrodes tend to vibrate mechanically, giving rise to effects characterized by the term *microphonic action.* In audio-frequency amplifiers these vibrations cause changes in the plate current that are of audible frequency; thus they are amplified along with the desired signal. In radio-frequency amplifiers the vibration of the tube structure will vary the amplification slightly, causing the microphonic noises to be modulated upon the radio-frequency signal.

Vibrations may be transmitted to the tube either mechanically through the tube socket or acoustically through the action of sound waves. Microphonic effects are most frequently encountered when tubes are in the immediate vicinity of loud-speakers or when equipment is in the presence of intense vibrations, as on airplanes, etc., and they are particularly troublesome in high-gain audio-frequency amplifiers. Microphonic noises can be reduced by mounting the tubes on spring suspensions and protecting them from sound waves. In addition, different types of tubes and individual tubes of the same type vary greatly in their tendency toward microphonic action. Hence, when microphonic effects are present, the proper procedure is to try a number of different tubes of the same type in the first stage of the amplifier. If the best results obtained in this way are not satisfactory, special non-microphonic tubes designed to provide a very rigid structure should be used.

Noise (Thermal Agitation, Shot Effect, and Related Phenomena).—Noise is the name given to irregular sizzling, frying, or crackling sounds to which no definite pitch can be assigned. Noise in amplifiers may be caused by poor contacts, by faulty resistances, by failing condensers, or by exhausted batteries.

Carbon resistors carrying direct current are a particularly troublesome source of noise and so cannot be used as plate-coupling resistances when followed by more than one stage of audio-frequency amplification. This noise arises from fluctuations in the contact resistance between adjacent granules and is similar in character to the "hiss" occurring in carbon microphones.[2] The value of the noise voltage is roughly proportional to the product $IR^{1.6}$, where I is the current and R is the resistance through which the current flows.

After all other sources of noise have been eliminated, there is still a residual output produced by thermal agitation of electrons, shot effect,

[1] For further information see: Alan C. Rockwood and Warren R. Ferris, Microphonic Improvements in Vacuum Tubes, *Proc. I.R.E.*, vol. 17, p. 1621, September, 1929 D. B. Penick, The Measurement and Reduction of Microphonic Noises in Vacuum Tubes, *Bell System Tech. Jour.*, vol. 13, p. 614, October, 1934.

[2] See C. J. Christensen and G. L. Pearson, Spontaneous Resistance Fluctuations in Carbon Microphones and Other Granular Resistances, *Bell System Tech. Jour.*, vol. 15, p. 197, April, 1936.

and related phenomena.[1] The noise these sources produce is a sizzling or frying sound representing energy more or less uniformly distributed over the entire frequency spectrum.

The most important source of noise in properly operated amplifiers arises from thermal agitation of the electrons of the conductors in the input circuit of the amplifier. It is well known that the conductivity of metals is a result of the presence of free electrons, and that these electrons are continuously moving about in the conductor at a velocity that depends upon the temperature. At any one instant there will ordinarily be more electrons moving in one direction than in the other, causing a voltage to develop across the terminals of the conductor. This voltage will vary from instant to instant in an irregular manner in accordance with the predominant motion of the electrons in the conductor. The square of the voltage developed in this way across the terminals of an impedance is directly proportional to the resistance component of the impedance and also directly proportional to the absolute temperature of the impedance, while the energy is uniformly distributed over the entire frequency spectrum from zero frequency up to frequencies much higher than those used in radio communication. Thus the magnitude of the noise voltage produced on frequencies lying between 1000 and 2000 cycles is just the same as the magnitude of the noise voltage between 1,000,000 and 1,001,000 cycles.

The magnitude of the voltage produced by thermal agitation can be calculated by the formula

$$\left.\begin{array}{l}\text{Square of effective value of voltage com-}\\ \text{ponents lying between frequencies } f_1 \text{ and } f_2\end{array}\right\} = E^2 = 4kT\int_{f_1}^{f_2} R\,df \qquad (110)$$

where

k = Boltzmann's constant = 1.374×10^{-23} joule per degree Kelvin

T = absolute temperature in degrees Kelvin

R = resistance component of impedance producing voltages of thermal agitation (a function of frequency)

f = frequency.

In the special case where the resistance component of the impedance is constant over the range of frequencies from f_1 to f_2, Eq. (110) reduces to the much simpler form

$$E^2 = 4kTR(f_2 - f_1) \qquad (111)$$

[1] Further information on thermal agitation, shot effect, and related noises in vacuum-tube circuits is to be found in the following references:

F. B. Llewellyn, A Study of Noise in Vacuum Tubes and Attached Circuits, *Proc. I.R.E.*, vol. 18, p. 243, February, 1930.

J. B. Johnson and F. B. Llewellyn, Limitation to Amplification, *Elec. Eng.*, vol. 53, p. 1449, November, 1934.

G. L. Pearson, Fluctuation Noise in Vacuum Tubes, *Bell System Tech. Jour.*, vol. 13, p. 634, October, 1934.

The noise voltages arising from thermal agitation set an ultimate limit to the lowest potential that can be amplified without being lost in a background of noise. An idea as to the order of magnitude of this limit can be obtained by noting that with a resistance of ½ megohm at 300°K., the noise voltage according to Eq. (111) for a frequency band 5000 cycles wide is 6.4 μv. The noise level of most amplifiers is determined by the thermal agitation of the resistance in the grid circuit of the first amplifier tube. However, where this resistance is quite low, the thermal agitation arising in the plate resistance and plate load impedance of the first tube may become the limiting factor.[1]

The shot effect results from the fact that the stream of electrons flowing from cathode to plate is made up of a series of particles rather than a continuous fluid. As a result, the electron flow to the plate is somewhat irregular, resembling hailstones striking a metal surface, and this gives rise to slight irregularities in the plate current of the vacuum tube and hence to noises in the amplifier. The presence of a space charge in the vacuum tube tends to smooth out the irregularities in the arrival of electrons at the plate, and this smoothing effect is so great as practically to eliminate the shot effect when complete temperature saturation exists. *It is therefore very important that the electron emission from the cathode be sufficient to produce an adequate space charge if the noise level is to be kept low.* The irregularities produced by the shot effect represent a distribution of alternating-current energy that is substantially uniform throughout the frequency spectrum, just as is the case with thermal agitation noise. The magnitude of shot effect in the presence of an incomplete space charge is proportional to $\partial I/\partial J$, where I is the total space current and J is the total electron emission from the cathode. The adequacy of the space charge from the point of view of eliminating shot effect can therefore be determined by increasing the cathode temperature and noting the extent to which the space current increases.[2]

There are two other important sources of noise present in tubes. The first of these is a form of shot effect called *flicker effect;* it is caused by

[1] The thermal agitation voltage E appearing across a plate load resistance R_0 as a result of thermal agitation in R_0 and in the plate resistance R_p of the tube is given by the relation

$$E^2 = 4k\frac{R_pR_0}{(R_p + R_0)^2}(T_0R_p + T_fR_0)(f_2 - f_1) \tag{112}$$

where T_0 and T_f are the temperatures of load resistance and cathode of the tube, respectively. Note that the thermal agitation in the plate resistance corresponds to the temperature of the hot cathode.

[2] For formulas from which one can calculate the noise arising from shot effect, see G. L. Pearson, *loc. cit.*, or F. B. Llewellyn, *loc. cit.*

changes of emission over small cathode areas. Flicker effect is relatively large in tubes employing oxide-coated filaments, where it normally overshadows the true shot effect. The second source of noise arises from positive ions produced as a result of ionization of residual gas, or as a result of positive-ion emission by the cathode (particularly common in tungsten filaments). Such positive ions cause a shot effect because they upset the space-charge equilibrium. With tubes having a good vacuum, the noise introduced in this way is of the same order of magnitude as the thermal agitation noise occurring in the plate circuit of the tube.

50. Input Admittance of Amplifier Tubes.—The input admittance of a tube is the admittance that is observed between the grid and cathode terminals when looking toward the tube. This input admittance takes into account the current that flows into the capacity between the

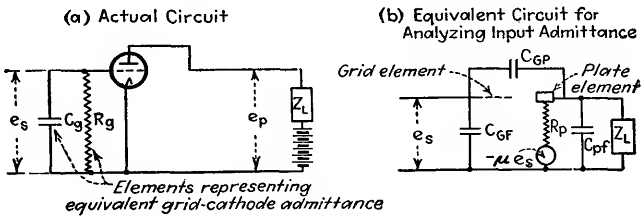


FIG. 130.—Equivalent circuit for analyzing amplifier input admittance.

grid electrode and grounded electrodes such as the cathode and screen, and also includes the effect of whatever current flows to the plate electrode as a result of the capacity between plate and grid and the potential difference between these electrodes. This latter component of the grid current depends upon the load impedance in the plate circuit because the alternating-current voltage between plate and grid is the difference between the signal voltage applied between grid and cathode and the amplified voltage developed across the plate load impedance, and the latter obviously depends upon the load. When the plate load impedance is great enough to produce appreciable amplification, the potential difference between grid and plate electrodes will be considerably greater than the signal voltage, with the result that a relatively large grid current flows from grid to plate, causing this part of the input admittance of the tube to be very important. If the load impedance in the plate circuit is a resistance, the input admittance of the tube is a pure capacity, but, if the plate load impedance has a reactive component, the input admittance of the tube will have a resistance component even though the grid is at a negative potential with respect to the cathode and attracts no electrons.

The input admittance of a vacuum tube can be represented by a resistance in parallel with a condenser, as shown in Fig. 130. If the

ratio of the amplified voltage developed across the load impedance in the plate circuit to the signal voltage being amplified is called A, and θ is used to represent the difference in phase between the voltage across the load impedance and the equivalent voltage $-\mu e_s$ acting in the plate circuit, then the resistance and capacity components of the input admittance of an amplifier tube are given by the following formulas:[1]

$$\text{Input resistance} = R_g = -\frac{1/\omega C_{gp}}{A \sin \theta} \tag{113}$$

$$\text{Input capacity} = C_g = C_{gf} + C_{gp}(1 + A \cos \theta) \tag{114}$$

where

C_{gp} = grid-plate tube capacity
C_{gf} = grid-cathode tube capacity
A = ratio of voltage developed across load impedance in plate circuit to applied signal (*i.e.*, A is the amplification of tube alone, not taking into account any step-up of voltage in the load)
θ = angle by which voltage across load impedance leads equivalent voltage acting in plate circuit (θ positive for inductive load impedance).

Examination of Eq. (114) shows that, with zero plate load impedance, the input capacity has the value $(C_{gf} + C_{gp})$ and reaches a value of $(C_{gf} + (1 + \mu)C_{gp})$ at very high load impedances (provided no negative resistances are present in the plate circuit). It is to be noted that the

[1] The derivation of Eqs. (113) and (114) is as follows: Referring to Fig. 130, E_s will be used to represent the signal voltage applied to the grid, and E_p the magnitude of the amplified voltage developed across the load impedance between anode and cathode. The voltage E_p leads $-\mu E_s$ by an angle θ and hence leads E_s by an angle $(\theta + 180°)$. With these definitions the voltage across the grid-plate tube capacity C_{gp} is $(E_s - E_p\underline{/\theta + 180°})$, and the current flowing from grid to plate as a result of this voltage across C_{gp} is $j\omega C_{gp}(E_s - E_p\underline{/\theta + 180°})$. The current flowing from grid to cathode through the grid capacity C_{gf} is $j\omega C_{gf}E_s$, so that the total grid current, which is the sum of these, is

$$\begin{aligned}\text{Total grid current} &= j\omega C_{gf}E_s + j\omega C_{gp}(E_s - E_p\underline{/\theta + 180°})\\ &= \omega C_{gf}E_s\underline{/90°} + \omega C_{gp}(E_s\underline{/90°} - E_p\underline{/\theta + 270°})\end{aligned}$$

This total current divided by the voltage E_s gives the admittance of the grid, which is therefore

$$\text{Admittance of grid} = \omega C_{gf}\underline{/90°} + \omega C_{gp}\left(1\underline{/90°} - \frac{E_p}{E_s}\underline{/\theta + 270°}\right)$$

The real part of this admittance represents the input conductance (*i.e.*, the reciprocal of the input resistance), while the quadrature part is the input susceptance, which when divided by ω gives the input capacity. Equations (113) and (114) are merely these two components of the input admittance with E_p/E_s denoted by the symbol A.

input capacity of the tube is dependent only upon the amplification A, the phase shift θ, and the tube capacities, and is independent of frequency; although at high frequencies the same input capacity is more important because it draws a greater current.

The input resistance of the vacuum tube may be either positive or negative, as seen from Eq. (113). A positive input resistance results when the load impedance in the plate circuit is a capacitive reactance, while a negative resistance is obtained with an inductive load in the plate circuit. *A positive input resistance means that energy is transferred from the grid to the plate through the grid-plate capacity, while a negative input resistance indicates that the phase relations are such that energy is transferred from the output or plate circuit of the tube to the grid circuit.* The value of input resistance for the same amplification A and phase shift θ varies inversely as the frequency and may be very low at high frequencies.

Special attention must be given to the input admittance of triodes, since here the grid-plate capacity C_{gp} is large. With screen-grid and pentode tubes the direct capacity between grid and plate is so small as to be practically zero, and the input admittance with such tubes can generally be considered as simply the grid-cathode capacity plus the grid-screen capacity.

Effects of Input Admittance in Audio-frequency Amplifiers.—In audio-frequency amplifiers the capacity component of the input admittance is of importance because it usually represents the major part of the capacity shunting the amplifier, and hence is the principal factor determining the response at high frequencies. The input resistance is normally not important because it is very high at audio frequencies.

In pentode and screen-grid tubes the input capacity is for all practical purposes equal to the sum of the grid-cathode and grid-screen capacities, and is not affected by the impedance in the plate circuit of the tube. With triode tubes, however, the input capacity tends to be large because of the large capacity between grid and plate and the large potential difference that exists across this grid-plate capacity as a result of the amplification A in the tube. The effective input capacity is then roughly proportional to the amplification factor of the tube, and becomes extremely large with high-mu tubes. This is one of the important reasons why triode tubes, particularly high-mu triode tubes, are not so suitable as pentode tubes for voltage amplifiers.

Effects of Input Admittance with Tuned Amplifiers.—When the plate load impedance is a tuned circuit, the magnitude A and the phase shift θ of the amplification vary greatly with frequency, and hence the part of the input resistance and capacity resulting from capacity between the grid and plate electrodes will go through corresponding changes.

The variations of input resistance and capacity in a typical tuned radio-frequency amplifier are shown in Fig. 131. The input capacity curve is similar in character to the amplification curve but is more peaked. The input resistance goes through wide variations, being positive for frequencies higher than resonance (plate load impedance a capacitive reactance), negative for frequencies below resonance (plate load impedance an inductive reactance), and infinite at resonance. The value of grid-plate capacity assumed in Fig. 131 is small and might readily occur in a pentode tube if the utmost care is not used to avoid all possible direct coupling between grid and plate circuits. If the capacity is larger

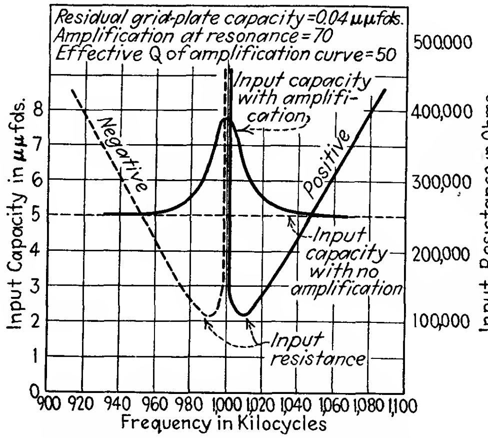


FIG. 131.—Curves of input capacity and input resistance of a tuned amplifier having a small but not negligible grid-plate capacity.

or if the amplifier gain is greater, the effects will be even more pronounced. When a tuned circuit is connected between the grid and cathode of a tube having an input impedance such as shown in Fig. 131, the resonance characteristics of this tuned circuit are considerably altered. Thus, when the tuned circuit connected between grid and cathode has a lower resonant frequency than the tuned circuit forming the plate load impedance, the negative input resistance of the tube neutralizes at least some of the resistance of the resonant circuit associated with the grid; if the negative resistance is less than the parallel resonant impedance of the tuned circuit across the grid, oscillations will result. When the resonant circuit connected across the grid is tuned to the same frequency as the resonant circuit in the plate, then the resonance curve of the grid tuned circuit can be expected to be seriously distorted as a result of the fact that upon one side of resonance the tube presents a positive resistance and on the other side a negative resistance. This is illustrated in Fig. 132.

Neutralization of Input Admittance of Vacuum-tube Amplifiers.—The effects that are produced by the transfer of energy between the grid and plate circuits of a vacuum-tube amplifier through the grid-plate tube capacity can be neutralized by an electrical network that transfers an equal amount of energy in the opposite direction. There are a number of ways in which this operation can be carried out, the most common of which are shown in Fig. 133. All these arrangements employ a neutraliz-

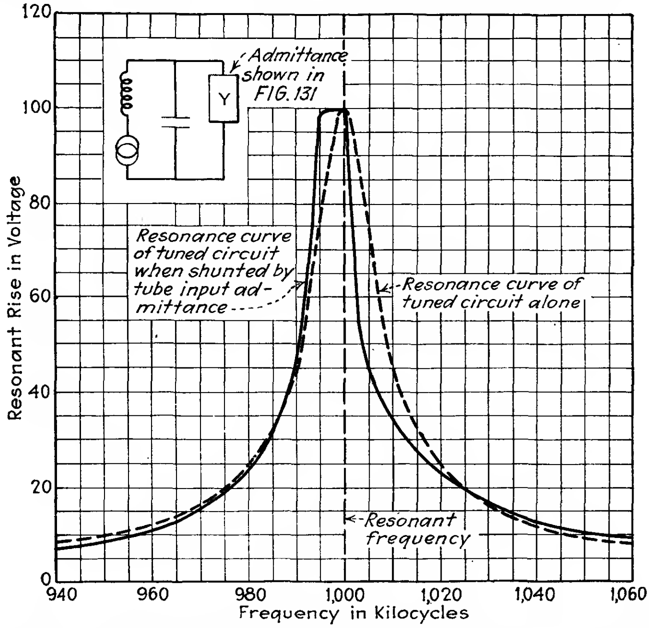


FIG. 132.—Unsymmetrical resonance curve obtained when a resonant circuit is shunted by an admittance similar to that of Fig. 131.

ing condenser C_N connecting the input (*i.e.*, grid) circuit and the output (*i.e.*, plate) circuit in such a way that the current passing through the neutralizing condenser is of the proper amplitude and phase to neutralize exactly the effect of the current flowing between plate and grid circuits of the amplifier *via* the grid-plate tube capacity. Consider the circuit of Fig. 133*a*, which consists of an ordinary transformer-coupled radio-frequency amplifier to which there has been added a neutralizing inductance L_N connected in such a way that the voltage at the end of this coil connected to the neutralizing condenser C_N is in phase opposition to the voltage at the plate end of the primary inductance L_p. The voltage across L_N is then applied to the neutralizing condenser and causes the grid to receive a current that, with proper size of C_N, is equal in magnitude and opposite in phase to the current flowing through the grid-plate tube capacity, and so completely neutralizes the energy transfer through the

tube capacity. The input capacity of such an amplifier with perfect neutralization is $C_{gf} + C_{gp} + C_N$, and the input resistance is infinite. If the coupling between L_N and L_p is very close, the neutralization is substantially independent of frequency.

A neutralizing circuit of a slightly different type is shown at Fig. 133*b*. In this arrangement the neutralizing condenser C_N is given such a capacity that the current through it as a result of the voltage developed in the plate circuit is equal in magnitude to the current passing through the grid-plate tube capacity; but since these two currents produce effects in the tuned circuit that are in phase opposition, they neutralize each other. This neutralization is theoretically independent of the frequency

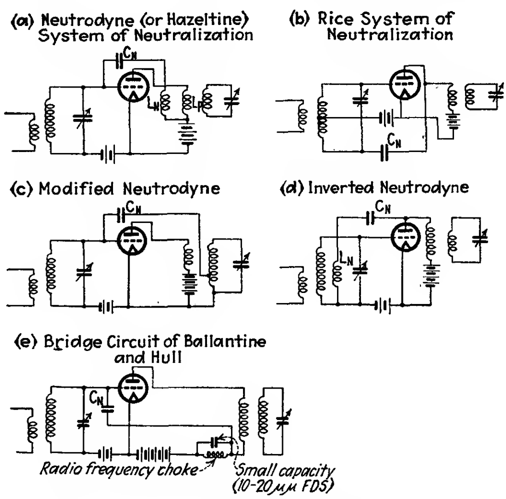


FIG. 133.—Typical neutralizing circuits.

at which the tuned circuit is resonant and under practical conditions can be made approximately so. Several additional types of neutralizing circuits are shown in Fig. 133, and still other arrangements have been devised; but since all these make use of the same general principles that are involved in the two specific cases discussed above, they need not be given special consideration.

A different and sometimes very profitable viewpoint is to consider the neutralizing arrangement as a bridge in which the output and input circuits are connected across the opposite diagonals. When the bridge is balanced, the input circuit receives no energy from the output circuit because the two are in electrically neutral locations with respect to each other.

Perfect neutralization cannot be maintained in practice over a wide band of frequencies because leakage inductances and stray capacities prevent the neutralizing current from being exactly proportional to, and

exactly out of phase with, the current through the grid-plate tube capacity at all frequencies. Imperfect neutralization gives rise to a certain amount of energy transfer, but it is possible to maintain the balance sufficiently well under actual operating conditions to make neutralized tuned-triode amplifiers operate satisfactorily. Figure 134 shows the effect that various degrees of unbalance in the neutralization system produce upon the input capacity and resistance of a typical tuned radio-frequency

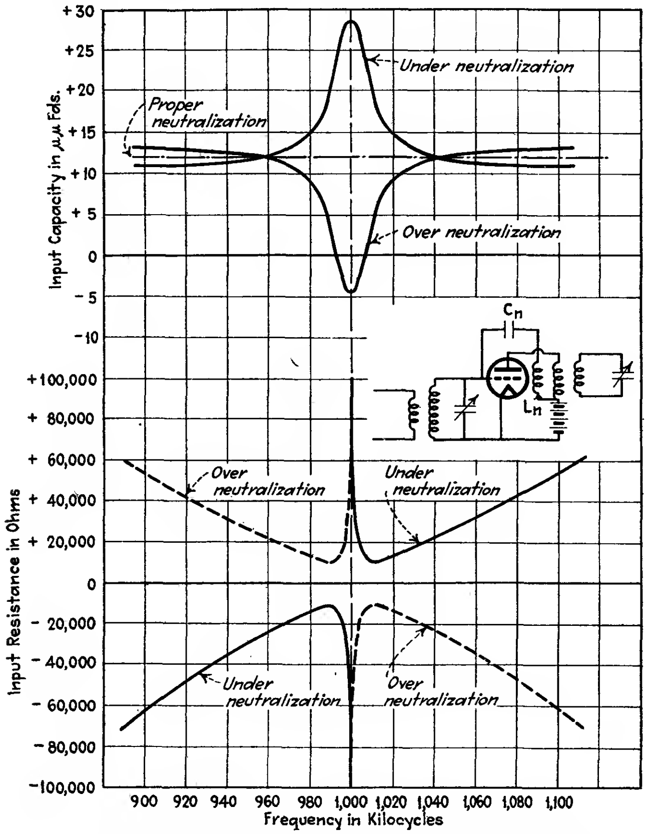


FIG. 134.—Curves showing the effect of improper neutralizing capacity C_n on input resistance and capacity of a tuned radio-frequency amplifier.

amplifier. With perfect balance the input resistance is infinite and the input capacity is constant. With insufficient neutralization the input resistance and capacity curves are similar to those for no neutralization but vary through a smaller range, while overneutralization causes the input capacity to be negative near resonance and makes the input resistance negative at high frequencies and positive at low frequencies. The practical effects of overneutralization are thus similar to those of underneutralization except that the distortion in the input resonant circuit is

of opposite symmetry from that with insufficient neutralization. In either case the effect is to distort the resonance curve of the input circuit in a manner analogous to that shown in Fig. 132.

Neutralization is always necessary in tuned triode amplifiers for otherwise the input resistance will be so low that oscillations can be expected. It is not employed with pentode and screen-grid amplifiers, however, because the direct-capacity coupling between grid and plate in such tubes is very small. If trouble is encountered from this residual grid-plate capacity, the usual procedure is to reduce the amplification per stage. An alternative possibility is to place a resistance of a few thousand ohms in series with the grid of the tube as shown in Fig. 135. Such a resistance introduces proportionately more loss at the higher frequencies because of the increased current flowing to the grid-screen and grid-cathode capacities, and thereby tends to counteract the lower input admittance as the frequency is increased.

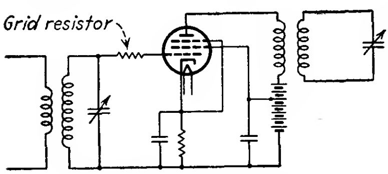


FIG. 135.—Tuned amplifier with grid resistor to reduce the tendency toward oscillation as a result of the residual grid-plate capacity.

Neutralization is seldom employed in audio-frequency amplifiers even with triodes because at low frequencies one need consider only the input capacity, and by proper design this can be taken care of without neutralization. It is possible, however, to improve the high-frequency response of a given triode amplifier by neutralizing to reduce the input capacity, and in some cases there is an advantage in overneutralizing.[1]

Input Admittance at Ultra-high Frequencies.—At very high frequencies the length of time it takes an electron to travel from the cathode to plate can no longer be considered as negligible compared with the length of time represented by a cycle. When this situation exists, it is found that there is an additional input loss as a result of energy that is supplied by the grid electrode to the electrons traveling toward the plate. This transfer of energy takes place even when the grid is negative and attracts no electrons to itself. The input resistance resulting from this is inversely proportional to the square of the frequency, and it becomes one of the limiting features of tube operation at frequencies exceeding about 20 mc. The phenomenon is discussed further in Sec. 53.

51. Multistage Amplifiers with Special Reference to Regeneration.—Under ideal conditions the gain of a multistage amplifier is the product of the amplification of the individual stages, and the frequency response is therefore the product of the frequency-response characteristics of the individual stages.

[1] See Paul W. Klipsch, Applying Neutralization to Audio-frequency Amplifiers, *Electronics*, vol. 7, p. 252, August, 1934.

It is generally found, however, that the total amplification of an actual multistage amplifier is not exactly equal to the product of the amplifications developed by the individual stages when acting as separate single-stage amplifiers. This is because of *regeneration*, which is the name given to the effects produced by the transfer of energy between stages. This transfer can take place through stray electrostatic capacity existing between the grid and plate electrodes of the amplifier tubes, or as a result of impedances common to two or more stages of amplification.

Because of the large differences in power level encountered in multistage amplifiers as a result of the fact that the energy level in the output stages is enormously greater than the energy level in the first stages of a high-gain amplifier, it takes very little coupling between the output and input of the amplifier to have considerable effect. Thus, if the total voltage amplification is 100,000, the ratio of powers at input and output points, assuming the same load resistances, is 10^{10}, and it is obvious that even a very small portion of the output energy transferred back to the input will be large in comparison with the input signal power that is being amplified. It is further apparent that the effect of stray couplings will be greater the higher the gain of the amplifier.

When the regeneration is large, the amplifier will usually oscillate and so become unusable. Even if the regeneration is not sufficient or of the proper phase to produce oscillations, it will normally change the amplification and alter seriously the frequency-response characteristic; thus it is to be avoided.

Regeneration in Audio-frequency Amplifiers.—The most common source of regeneration in audio-frequency amplifiers arises from plate impedances common to two or more stages, but other factors, such as stray magnetic and electrostatic couplings and the transfer of energy through the grid-plate capacity of the tubes, must be considered, especially if the total amplification is very high. The principal source of magnetic coupling in audio-frequency amplifiers is between transformers, particularly between the input and output transformers of the amplifier. The remedy is to avoid transformers wherever possible by the use of resistance coupling and, where transformers must be employed, to use shielded transformers spaced as far as possible and oriented in such a way as to minimize coupling. Trouble from magnetic coupling is also occasionally caused by improper arrangement of wiring.

Electrostatic coupling in audio-frequency amplifiers is readily eliminated by shielding. It is usually sufficient to provide each tube with an electrostatic shield, although, sometimes, completely inclosing one or more low-level stages is desirable.

The input admittance of the tube modifies the characteristics of the amplifier at high frequencies. With pentodes the input admittance is

essentially a capacity equal to the sum of the grid-cathode and grid-screen electrostatic capacities. In the case of triodes the input admittance depends upon the load impedance in the plate circuit of the tube (see Sec. 50), and this impedance in turn depends upon the input admittance of the next stage. The effect of the input admittance in such cases can be accurately taken into account by computing the amplification stage by stage, beginning with the output tube and working backward toward the input. In doing this the amplification of each stage is calculated by noting that the input admittance of each tube is shunted across the load of the preceding tube and by following the usual procedure for amplifier calculations.

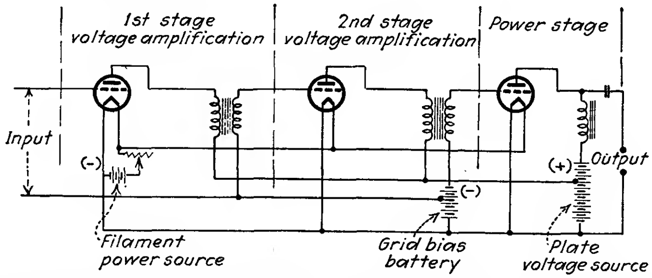


FIG. 136.—Diagram of a multistage audio-frequency amplifier, consisting of two stages of voltage amplification followed by a power stage, with a common source of plate impedance.

Regeneration from a common plate impedance occurs where the different stages receive their plate voltage from a common source, as is the case in Fig. 136, since the internal impedance of this voltage source is common to the different stages and hence provides a coupling common to the plate circuits of the tubes involved. This results in regeneration that may either increase or decrease the amplification, the exact effect depending on the phase relations. A common plate impedance representing the internal impedance of a source of plate voltage is usually the most important cause of regeneration in a carefully built audio-frequency amplifier, and it produces effects that are of such importance as to warrant detailed consideration.

When the plate current of more than one amplifier tube flows through the same common impedance, the voltage drop which the currents from one stage develop across the common impedance will transfer energy to all the other stages. The part of this action that is important in determining the characteristics of a multistage amplifier is the transfer of energy from the plate circuit of the last stage of amplification to the plate circuit of the first stage. The energy transferred between plate circuits of other than the last and first tubes is relatively unimportant because the difference in energy level between the end tubes is much greater than that between any other pair. The result is that the effect of a common plate-circuit impedance on the behavior of a multistage

amplifier can be analyzed with a high degree of accuracy by considering that the impedance of the voltage source is common to only the first and last tubes fed from this source. This procedure simplifies the problem, making it possible to predict with accuracy when trouble from regeneration can be expected and permitting simple quantitative calculation of the magnitude of the effects that will be produced.

Regeneration from a common plate impedance is characterized by the following two fundamental principles: first, the principal effect of the

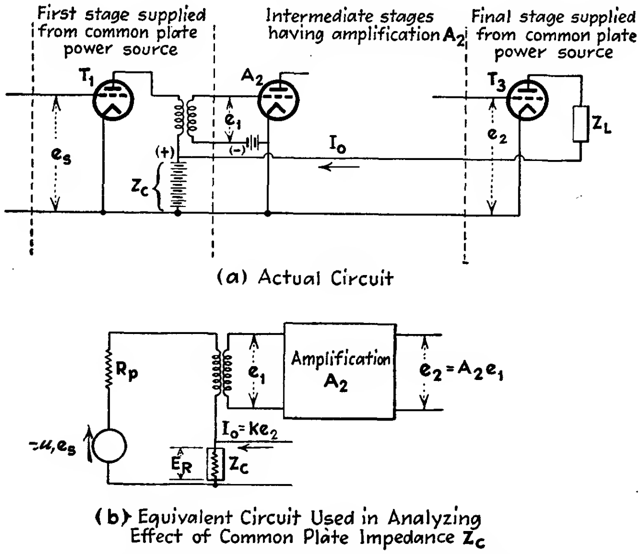


FIG. 137.—Circuit of multi-stage transformer-coupled amplifier having a common plate impedance resulting from the use of a common source of plate power, together with equivalent circuit that can be utilized to analyze the effect of the common plate impedance

regeneration is equivalent to altering the gain of the first stage of the amplifier, while leaving the remainder of the amplifier unchanged; second, *the effects produced by the regeneration become important only when the voltage that is coupled or otherwise transferred back into the plate circuit of the first tube is of appreciable magnitude compared with the signal voltages present in this circuit in the absence of regeneration.*

The detailed analysis of the effect of a common plate impedance is carried out by first setting up equivalent circuits, such as done in Figs. 137 and 138, where the first stage consists of a transformer-coupled triode and resistance-coupled triode, respectively.[1] Here the voltage e_2 acting at the grid of the final tube produces a plate current that causes

[1] Much of the material in this discussion of the quantitative effects produced by a common plate impedance represents the results of a research carried out at Stanford University under the supervision of the author by D. H. Ring. Among other things, this study showed that the effect of a common plate impedance could be accurately calculated by Eqs. (115) and (116).

a voltage drop E_R in the common plate impedance. This voltage E_R acts in the plate circuit of the first amplifier tube to produce currents that modify the voltage which the first tube applies to the second tube of the

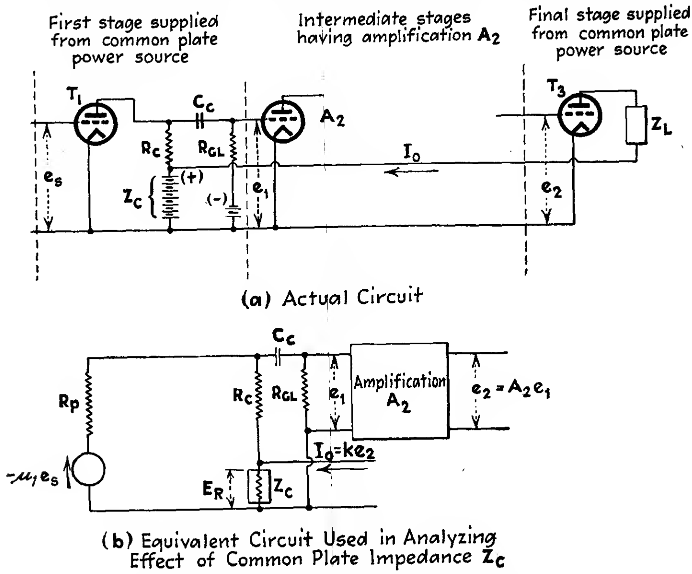


FIG. 138.—Circuit of multi-stage resistance-coupled amplifier having a common source of plate power, and equivalent circuit for analyzing the effect which the common plate impedance has upon the amplification.

amplifier. The result is therefore equivalent to modifying the gain of the first stage. Analysis of the equivalent circuits leads to the results:[1]

$$\left.\begin{matrix}\text{Actual amplification of}\\ \text{first stage having trans-}\\ \text{former coupling}\end{matrix}\right\} = \frac{\left\{\begin{matrix}\text{amplification that would}\\ \text{exist without feed back}\end{matrix}\right\}}{1 - \dfrac{A_1A_2KZ_c}{\mu_1}} \tag{115}$$

[1] Equation (115) is derived by noting in Fig. 137a that the total voltage acting in series with the plate circuit of the first tube is $-\mu e_s + E_R$. Hence

$$e_1 = A_1\left(-e_s + \frac{E_R}{\mu_1}\right)$$

But

$$E_R = e_2KZ_c = A_2e_1KZ_c$$

Hence

$$e_1 = A_1\left(-e_s + \frac{A_2e_1KZ_c}{\mu_1}\right)$$

Solving this for the amplification $|e_1/e_s|$ of the first stage gives Eq. (115).

Equation (116) is derived in a somewhat similar manner by noting that, since the voltage E_R acts in series with the coupling resistance R_c, it is equivalent to a constant-current generator of current E_R/R_c (just as a voltage $-\mu e_s$ in series with a resistance R_p is equivalent to a constant-current generator $-\mu e_s/R_p = -G_m e_s$).

$$\left.\begin{array}{l}\text{Actual amplification of}\\ \text{first stage having re-}\\ \text{sistance coupling}\end{array}\right\} = \frac{\left\{\begin{array}{l}\text{amplification that would}\\ \text{exist without feed back}\end{array}\right\}}{1 - \dfrac{A_1A_2KZ_c}{G_mR_c}} \tag{116}$$

where

A_1 = vector amplification of first stage when there is no common impedance (does not include phase reversal produced by tube).

A_2 = vector amplification from output of first tube to grid of last tube

Z_c = common plate impedance (a vector having magnitude and phase)

K = factor which when multiplied by voltage on grid of last tube gives the a-c current through the common plate impedance ($E_R = I_0Z_c = e_2KZ_c$). This factor depends on the constants of last tube and the impedance in its plate circuit

μ_1 = amplification factor of first tube

G_m = mutual conductance of first tube

R_c = coupling resistance in resistance-coupled case.

Consideration of the above equations shows that the modifying effect which regeneration has on the amplification increases as the over-all gain A_1A_2 of the amplifier becomes larger and as the common impedance Z_c is increased, *and is negligible only when the term on the right-hand side of the denominator of Eqs.* (115) *and* (116) *is small compared with unity.*

Regeneration troubles caused by a common plate impedance can be avoided by keeping the impedance as low as possible and by using filter circuits to isolate the input and output stages.[1] When batteries are used for supplying anode power, the common impedance consists of the internal resistance of the batteries shunted by whatever by-pass condenser is connected across the battery terminals, and the impedance will increase as the batteries become exhausted. With rectifier-filter systems the common impedance at frequencies above a few cycles per second is the reactance of the capacity shunting the output of the filter system. The common impedance then is negligible at moderate and high audio frequencies, but becomes high at very low frequencies. When the impedance of the power source is enough to modify the amplification appreciably, it is necessary to isolate the first and last stages by filters. These are normally located between the common plate impedance and the input stage, and consist of one or more resistances or inductances in series with the plate circuit, together with a corresponding number of capacity shunts to ground, as shown in Fig. 139. The proportions must be such that the series impedances are high compared with the shunt impedances

[1] The use of a push-pull power amplifier arrangement is also very helpful (see Sec. 59).

at all frequencies for which the filter is to be effective. The high series impedances then prevent the voltage developed across the common impedance from sending appreciable current toward the input stage, while the low shunt impedances short-circuit substantially all this

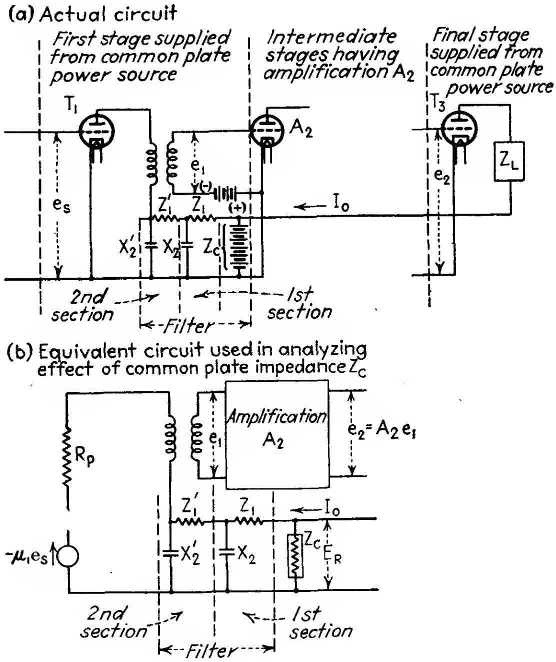


FIG. 139.—Diagram of the amplifier of Fig. 137 after filters have been added to minimize the effects of common plate impedance, together with equivalent circuit for analyzing the effect of energy transfer from the last stage to first stage through the common plate impedance when filters are present.

current. The result is that the voltage actually introduced into the plate circuit of the input tube is very much less than the voltage developed across the common impedance. The reduction is given approximately by the equations[1]

[1] These equations assume that X_2 is small compared with Z_1 and the plate and coupling impedances of the input tube, and that Z_1 is much larger than the common impedance. Under these conditions a voltage E_R across the common impedance Z_c causes a current E_R/Z_1 to flow through Z_1 (see Fig. 139). Most of this current flows through X_2 and so produces a voltage E_RX_2/Z_1 across X_2. In the case of a one-stage filter, this is the voltage, actually inserted in the plate circuit of the input tube, and it is seen to lead to a reduction amounting to X_2/Z_1. In the case of a two-stage filter, there is a further reduction of X_2'/Z_1', giving a total reduction of X_2X_2'/Z_1Z_1', as in Eq. (117b).

$$\left.\begin{array}{l}\text{Reduction with one-}\\\text{stage filter}\end{array}\right\} = \frac{X_2}{Z_1} \tag{117a}$$

$$\left.\begin{array}{l}\text{Reduction with two-}\\\text{stage filter}\end{array}\right\} = \frac{X_2 X_2'}{Z_1 Z_1'} \tag{117b}$$

where Z_1 and Z_1' are the series impedances of the filter sections and X_2 and X_2' are the reactances of the shunting condensers.

Motorboating in Audio-frequency Amplifiers.—In high-gain audio-frequency amplifiers having a good low-frequency response, regeneration from a common plate impedance often causes the amplifier to oscillate at a frequency of a few cycles per second. This is termed *motorboating* (from the "put-put" sound it causes in a loud-speaker), and is often difficult to eliminate because the shunt condensers of a filter become ineffective at very low frequencies.

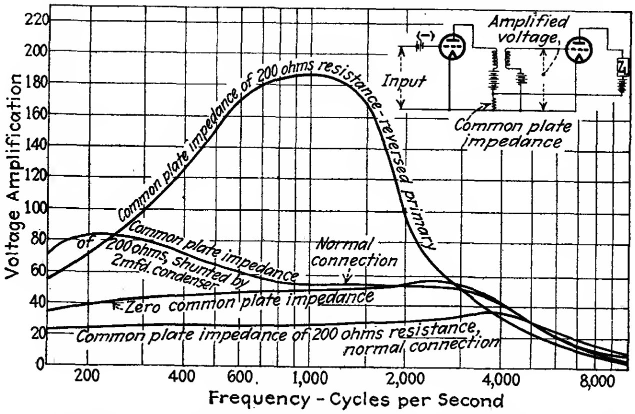


FIG. 140.—Experimental curves showing the effect of a common impedance in the plate circuit on the amplification characteristic of a transformer-coupled amplifier. The effect of the regeneration is to either increase or decrease the amplification according to the polarity of the transformer secondary. Shunting the common impedance by a large condenser eliminates the regeneration at high frequencies while allowing large regeneration to exist at low frequencies.

The best remedy for motorboating troubles is to make the low-frequency response no better than absolutely necessary, since by reducing the amplification at low frequencies the amount of regeneration and hence the tendency to oscillate is reduced. The common plate impedance should also be reduced to the lowest possible value by using high-capacity filter and by-pass condensers. When rectifier-filter power systems are employed, it is usually helpful to employ totally different smoothing filters for the input and output stages, together with low-impedance rectifier tubes and power transformers having low reactance. As a final resort, it is always possible to use a separate power-supply system for the input stage, and this is necessary if the over-all gain is very large.

In multistage audio-frequency amplifiers the mere absence of oscillation will not insure that the amplifier is operating properly. This is because regeneration insufficient to cause oscillations, or of the wrong phase to produce oscillations, may affect greatly the frequency response of the amplifier and may also alter the gain. This is illustrated by Figs. 140 and 141, which present experimental curves showing typical effects that can be produced by regeneration from a common plate impedance.

Regeneration in Multistage Tuned Radio-frequency Amplifiers.—The possible sources of regeneration are the same at radio as at audio

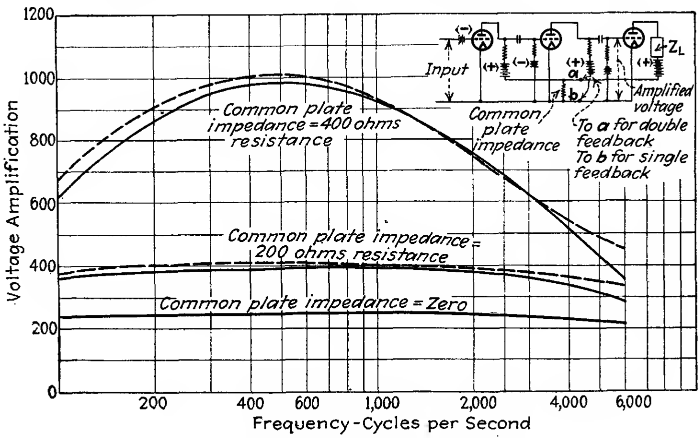


FIG. 141.—Experimental curves showing the effect of a common impedance in the plate circuit on the amplification characteristic of a three-stage resistance-coupled amplifier. The dotted curves are for the case where the impedance is common to only the first and last tubes while the solid curves are for the case where the impedance is common to all three plate circuits. The close agreement between the two is the practical justification for neglecting all energy transfer except that between first and last tubes when making calculations.

frequencies, but the precautions that must be used differ in the two frequency ranges.

Energy transferred from one part of a tuned amplifier to a preceding part may cause the amplification to be either greater or less than the ideal value, and the exact effect is equivalent to changing the resistance and the reactance of the tuned circuits in the amplifier. Such variations in reactance alter the resonant frequencies of the tuned circuits slightly and so are relatively unimportant, but the resistance changes that accompany regeneration have a profound effect. This is due to the fact that the actual resistance of tuned circuits is so low that only a few added ohms will largely destroy all resonance effects, while a few ohms subtracted from the actual circuit resistance can bring the effective resistance to zero, with the result that the amplification becomes infinite, *i.e.*, oscilla-

tions are produced. The character of the action in any particular case depends upon the phase and magnitude of the energy fed back to the amplifier input. Regenerative action is always greater as the amplification is increased, since then the extreme difference in the energy levels of different parts of the amplifier is larger. Regeneration also becomes greater the higher the frequency being amplified because capacities pass current and mutual inductances induce voltages in proportion to the frequency.

The magnitude of regeneration in ordinary radio-frequency amplifiers is usually sufficient to alter the characteristics completely unless special precautions are taken to eliminate the causes of energy feedback or to compensate for its effects. Regeneration resulting from stray capacitive and inductive couplings between different parts of the amplifier can be reduced by properly positioning these parts of the circuit with respect to each other, and can be eliminated completely by shielding with non-magnetic shields of the type described in Sec. 11. In orientating the various parts of the circuit, it is particularly important that the inductance coils of the different stages be as far apart as possible. It is also desirable to use coils of small dimensions so that the spacing between coils will be large compared with the coil size; this reduces the inductive coupling between coils. Electrostatic coupling between parts of the amplifier is not so difficult to control as inductive coupling between coils, and it can be kept small by proper arrangement of the wiring. It is impossible, however, to eliminate all electrostatic and magnetic coupling unless each stage of amplification is completely enclosed in a container made of copper or aluminum.

Impedances common to two or more stages are particularly troublesome causes of regeneration in radio-frequency amplifiers, because at the high frequencies involved a wire only a few inches long will often have sufficient reactance to provide an effective means of transferring energy. Unavoidable common impedances, such as those from common sources of electrode voltages, can be reduced to a low value by shunting with by-pass condensers, and in addition filters may be placed in the leads running to the individual stages. Indiscriminate use of the chassis as a return circuit often gives trouble since the chassis, when used in this way, provides an impedance common to all the circuits. It is therefore especially important that each coil of a resonant circuit be connected directly to the frame of the tuning condenser by a separate wire in order that the circulating current in the resonant circuit need not pass through the chassis.

Direct transfer of energy between the input and output circuits of the vacuum tube as a result of capacity between the grid and plate electrodes is kept small in tuned amplifiers by using pentode tubes in

which this direct capacity is very small (radio-frequency pentodes) and by shielding tubes and input and output circuits to prevent electrostatic couplings from existing outside the tube. With triodes, neutralization is always employed. By these means this source of regeneration is kept under control in tuned amplifiers, and, while always present to some extent, it does not seriously distort the amplifier behavior under most circumstances.

When regeneration in tuned amplifiers is appreciable, oscillations are usually produced, but, even if the regeneration is not sufficient to cause oscillations, it may distort the shape of the amplification characteristic, either by increasing the sharpness of resonance or by destroying the symmetry existing on either side of resonance, or both.

52. Feedback Amplifiers.[1]—In the feedback amplifier a certain amount of regeneration is deliberately introduced in such a way as to reduce the amplifier gain. By properly carrying out this operation it is possible to reduce the distortion and noise generated in the amplifier, to make the amplification substantially independent of electrode voltages and tube constants, and to reduce greatly the phase and frequency distortion.

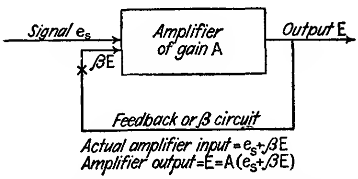


FIG. 142.—Schematic diagram of feedback amplifier.

The operation of a feedback amplifier can be understood by reference to the schematic diagram of Fig. 142. Here A represents an amplifier which has a gain A when used as an ordinary amplifier and which is supplied with a signal voltage e_s. Regeneration is introduced by superimposing on the amplifier input a fraction β of the output voltage E so that the actual input consists of a signal e_s plus the feedback voltage βE. The effective gain of the amplifier is then[2]

$$\left.\begin{matrix}\text{Gain, taking into}\\ \text{account feedback}\end{matrix}\right\} = \frac{A}{1 - A\beta} \tag{118}$$

In this equation the assumption as to signs is such that when the feedback voltage opposes the signal voltage, β is negative. The quantity $A\beta$ can be termed the feedback factor, and represents the amplitude of the voltage superimposed upon e_s compared with the actual voltage applied

[1] For further discussion see H. S. Black, Stabilized Feedback Amplifiers, *Elec. Eng.*, vol. 53, p. 114, January, 1934.

[2] Equation (118) is derived as follows: If E is the output voltage, then the feedback voltage is βE and the actual input potential is $(e_s + \beta E)$. This input amplified A times must equal E, *i.e.*,

$$(e_s + \beta E)A = E$$

Equation (118) then follows by solving for E/e_s, which is the actual gain.

to the input terminals. Thus if $A\beta = 50$, then for each millivolt existing between the input terminals the feedback voltage will be 50 mv, and, if the phase is such as to give negative feedback, a signal of 51 mv will be required to produce 1 mv at the amplifier input terminals.

Examination of Eq. (118) shows that, if the feedback factor $A\beta$ is large, the amplification is reduced by the presence of feedback, and furthermore, when $|A\beta| >> 1$, Eq. (118) reduces to

$$\left.\begin{array}{l}\text{Amplification with}\\ \text{large feedback}\end{array}\right\} = -\frac{1}{\beta} \tag{119}$$

Expressed in words, Eq. (119) states that, when the feedback factor $A\beta$ is large, the effective amplification depends only upon the fraction β of the output voltage that is superimposed upon the amplifier input, and *is substantially independent of the gain actually produced by the amplifier itself.*

This remarkable behavior is a result of the fact that, when the feedback is large, the voltage actually applied to the amplifier input terminals represents a small difference between relatively large signal and feedback voltages. A moderate change in the amplification A therefore produces a large change in the difference between signal and feedback voltages, thereby altering the actual input voltage in a manner that tends to correct for the alteration in amplification. Thus in the amplifier considered above where $A\beta = 50$, if the amplification A were halved by a change in design, it would then take 2 mv across the input terminals to deliver the same output as before. With β unchanged, the feedback voltage would still be 50 mv, so it would require a signal of $50 + 2 = 52$ mv instead of the previous 51 mv to produce the same output. Thus a 2 per cent change in effective over-all amplification results when the gain A is altered by 50 per cent.

Inasmuch as the quantity β depends upon circuit elements, such as resistances, that are permanent, the amplification with large feedback is substantially independent of the tube characteristics and electrode voltages. Furthermore Eq. (119) shows that the amplification with large feedback is inversely proportional to β, so that, if the fraction β of the output voltage that is superimposed upon the input is obtained by a resistance network, the amplification will be substantially independent of frequency and will have negligible phase shift. On the other hand, if it is desired to have the amplification vary with frequency in some particular way, this can be readily accomplished by making the β (or feedback) circuit have the same transmission-loss characteristic as the desired gain characteristic. This last property can be utilized in equalizing an amplifier.

The presence of negative feedback also greatly reduces the amplitude distortion produced in the amplifier. This distortion can be thought of as being generated in the amplifier, usually in the output stage. If d represents the amount of distortion appearing at the output in the absence of feedback when the output signal voltage is E, then in the presence of feedback the distortion with sufficient excitation to produce the same output voltage E will be less than d. This is because some of the distortion is fed back to the amplifier input through the feedback circuit and reamplified in such a way as to tend to cancel out the distortion originally generated. Represent by D the distortion voltages actually appearing in the output in the presence of feedback. The distortion voltage applied to the amplifier by the feedback circuit is then βD, and this is amplified A times by the amplifier. The total distortion output is then the distortion d actually generated in the amplifier plus the amplified feedback distortion βDA. That is, $D = d + \beta DA$, or

$$\left.\begin{array}{l}\text{Distortion with}\\ \text{feedback}\end{array}\right\} = D = \frac{d}{1 - A\beta} \tag{120}$$

$$= \frac{\text{Distortion in absence of feedback}}{1 - A\beta}$$

This shows that the amplitude distortion appearing in the output is reduced by the factor $(1 - A\beta)$. If $A\beta$ is made large by employing a large amount of feedback, the result is a very great reduction in the ratio of distortion to desired output.

The signal-to-noise ratio is also improved by feedback under certain conditions. A comparison of the signal-to-noise ratios in amplifiers with and without feedback, when the same noise voltage is introduced some place in the amplifier and the signal outputs are the same, shows that

$$\frac{\left.\begin{array}{l}\text{Signal-to-noise ratio}\\ \text{with feedback}\end{array}\right\}}{\left.\begin{array}{l}\text{Signal-to-noise ratio}\\ \text{without feedback}\end{array}\right\}} = \frac{a_f}{a_0(1 - A\beta)} \tag{121}$$

where a_f and a_0 are the amplification between the place at which the noise is introduced and the output, with and without feedback, respectively. Examination of this relation shows feedback will greatly reduce noises introduced in high-level parts of the amplifier, such as from a poorly filtered power supply in the plate circuit of the final tube. Feedback does not help reduce noise introduced at very low power levels, such as thermal agitation, induced hum, microphonic noises, etc., because the additional amplification required to make up for the loss in gain from

negative feedback also gives just enough extra amplification of the low-level noise to counteract the tendency of negative feed-back to neutralize the noise reaching the output.

The above discussion of the properties of feedback amplifiers can be summarized as follows for the condition where the feedback factor $A\beta$ is very much larger than unity: The amplification is very stable, the amount of amplification and the way in which it varies with frequency being substantially independent of the characteristics of the amplifier tubes, electrode voltages, or amplifier-circuit constants. The amount of gain, the way in which the gain varies with frequency, and the phase shift between input and output voltages can all be controlled by giving the β circuit the proper characteristics. In addition, the percentage of distortion generated in the output tube of the amplifier is reduced by the factor $(1 - A\beta)$, as is also all noise introduced into high-level portions of the amplifier. The price paid for obtaining these advantages is a reduction in gain, or, conversely, the addition of approximately one more stage of amplification than would otherwise be necessary to obtain a given net gain.

Feedback without Oscillations.[1]—In order to realize the advantages of feedback, the amplifier and its feedback must be so arranged that oscillations do not occur. This can be accomplished by arranging the circuits so that the feedback voltage is normally in phase opposition to the applied signal (*i.e.*, $A\beta$ negative and real) and by arranging the circuits so that there is no frequency where $A\beta$ is positive, real, and greater than unity.[2] The amplification and phase-shift characteristics of a feedback amplifier are therefore of fundamental importance in determining whether or not the amplifier will be stable.

With a single stage of resistance or impedance coupling, the amplification falls off at both high and low frequencies as shown in Fig. 95, which also gives the accompanying phase shift.[3] It will be noted that the maximum shift away from the normal mid-range value of zero is $\pm 90°$, and this occurs only at the extreme frequencies where the amplification

[1] Much of the material in this and the following divisions follows F. E. Terman, Feedback Amplifier Design, *Electronics*, vol. 10, p. 12, January, 1937.

[2] The exact criterion for avoiding oscillation in feedback circuits is that, when the value of $A\beta$ and its conjugate are plotted as a function of frequency on rectangular coordinates with the real part along the X-axis and the imaginary part along the Y-axis, the resulting curve will not inclose the point 1,0. This means that under some conditions oscillations will not occur even when $A\beta$ is positive, real, and greater than unity. See E. Peterson, J. G. Kreer, and L. A. Ware, Regeneration, Theory, and Experiment, *Proc. I.R.E.*, vol. 22, p. 1191, October, 1934.

[3] The phase-shift characteristics shown in Fig. 95 represent the phase shift with respect to the phase shift in the middle frequency range, and do not take into account the phase reversal produced in each stage as a result of the tube action.

drops to zero. At the point where the amplification has dropped to 70.7 per cent of its mid-range value, the phase shift is 45°, while at the 10 per cent point it is 84° 15′.

The characteristic of a transformer-coupled stage of voltage amplification is shown in Fig. 103. At low frequencies the behavior is identical with resistance coupling, with a maximum possible phase shift of 90° occurring at zero frequency where the amplification becomes zero. At high frequencies the situation is more complicated, however, as series resonance takes place between the leakage inductance of the transformer and the capacity in shunt with the secondary. At this series-resonant frequency, which is the frequency at which there will be a hump in the amplification characteristic if the plate resistance of the tube is too low, the phase shift is 90°, while at still higher frequencies the phase shift increases and reaches 180° at very high frequencies where the amplification is zero.

When a transformer is loaded with sufficient resistance across the secondary to prevent series resonance, as is the case with an output transformer, the amplification and phase-shift characteristics are then of the same character as with resistance coupling. The phase shift is zero in the mid-range of frequencies and reaches ±90° only at extreme frequencies where the output drops to zero (see Fig. 159).

The phase-shift characteristics of other types of amplifiers can be worked out as desired. In general, where the coupling arrangement employs a simple combination of resistance and reactance elements without series resonance, the maximum phase shift will be ±90° and this will occur only at extremes of frequency where the amplification becomes zero. However, in circuits where series resonance occurs, phase shifts up to ±180° are possible at extreme frequencies where the amplification becomes zero, and the amplification can be quite large with phase shifts exceeding 90°.

Phase shifts in the feedback network are important since these have as much effect on the amplifier as the phase of $A\beta$. Hence, if the feedback circuit is not a resistance network, its phase characteristics must be considered.

It is apparent from these phase-shift characteristics that, when a resistance feedback network is employed (*i.e.*, β with no phase shift), no trouble from oscillations need be expected when only a single-stage amplifier is involved. This is because, even with an interstage transformer, the phase cannot shift 180° from the normal value without the amplification and hence $A\beta$ dropping to zero. When only resistance, impedance, or output-transformer coupled stages are involved, oscillation troubles are not encountered even with two stages because, as the maximum possible phase shift per stage is 90°, the over-all phase will reach 180°

only at zero or infinite frequency, where the amplification and hence $A\beta$ are zero.

When more than two stages of amplification are involved, oscillations can always be expected unless the feedback factor $A\beta$ is small, or unless the amplifier is especially designed. Thus, if three identical resistance-coupled stages are involved, reference to Fig. 95 shows that the total phase shift will reach 180° (= 60° per stage) at frequencies where the amplification per stage has dropped to 0.50 of its mid-range value. Hence, to avoid oscillations with such a three-stage amplifier, the value of $A\beta$ in the mid-frequency range must be less than 8 in order to make $A\beta$ less than unity at frequencies where the phase shift becomes 180°. If greater values of feedback factor $A\beta$ are desired under such conditions, one of the stages must be designed to have a very wide frequency range so that this stage produces negligible phase shift compared with the remaining stages. Thus, if one desires $A\beta = 100$, then reference to Fig. 95 shows that at frequencies so far from the mid-range region that $A\beta$ has dropped to unity (*i.e.* frequencies for which the amplification per stage has dropped to 0.1 of the mid-range value) the total phase shift is $2 \times 84.25° = 168.5°$. Hence the remaining stage must not have a phase shift exceeding 11.5° for the same frequency range, and so must be designed with a correspondingly wide frequency range. These same principles can be applied to the design of amplifiers with four or more stages and to amplifiers having interstage transformers.

Practical Feedback Amplifiers.—A number of typical audio-frequency amplifier arrangements employing feedback are shown in Fig. 143.[1] In the circuits at *c*, *d*, *e*, and *f*, the feedback voltage is proportional to the output voltages and so causes the latter to be stabilized. In contrast with this, the circuits at *a*, *b*, and *g* make use of the current in the bias impedance of the output tube to develop the feedback voltage, so that, if the bias impedance is a resistance, the feedback action tends to stabilize the current in the output circuit rather than the voltage across it. Thus, when the output is a resistance-capacity network, the shunt capacity C_s causes the output voltage to drop off at high frequencies for constant a-c current in the plate circuit, and the coupling condenser C_c causes a similar falling off at low frequencies. Feedback action hence does not improve the frequency response of these circuits even though amplitude distortion is reduced. The falling off at high and low frequencies can, however, be greatly diminished by shunting the cathode resistor by a condenser C_2 to reduce the negative feedback at high frequencies, and by shunting by an inductance L_2 to reduce the negative feedback at low frequencies, as shown dotted in parts *a*, *b*, and *g*. By proper choice of these reactive

[1] A more detailed discussion of these circuits, together with descriptions of additional circuits is given by Terman, *ibid.*

elements the response can be made substantially flat over a much larger frequency range than would otherwise be possible. Circuits *a*, *b*, and *g* are not suitable when the output circuit contains an output transformer because there is no resulting improvement in frequency response, and the stabilization tends to make the magnetizing current of the transformer sinusoidal and so actually produces amplitude distortion in the output voltage.

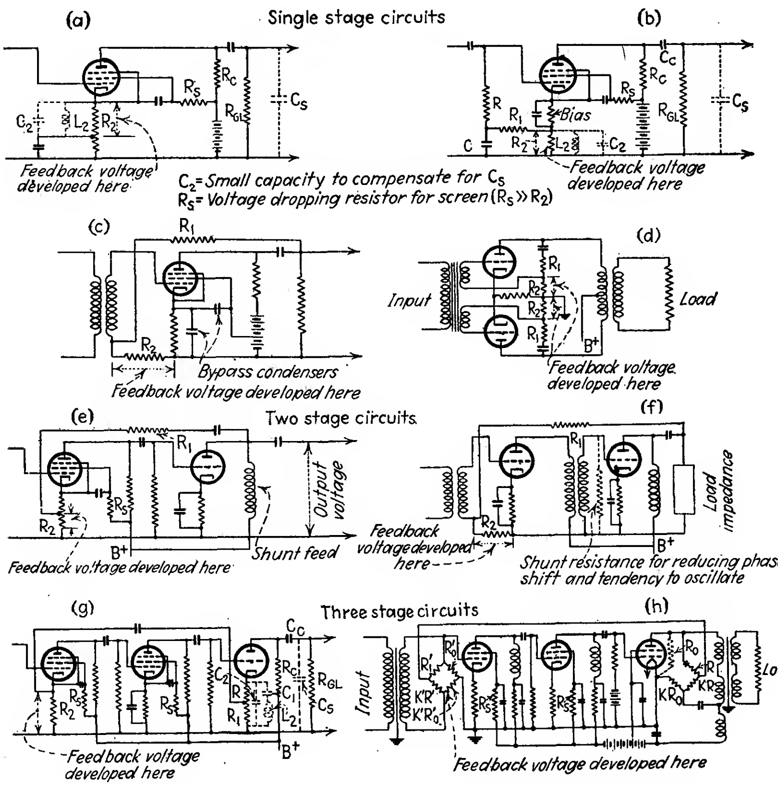


FIG. 143.—Circuit diagrams of typical feed back amplifiers.

When the bias resistor of pentode tubes is used for feedback purposes, it is essential that the suppressor grid be connected directly to cathode instead of to ground, that the screen potential be supplied through a series resistor, and that the screen be by-passed to the cathode by a condenser. The by-passing to the cathode is necessary, for otherwise the feedback voltage will be applied to the suppressor and screen, while the series resistor in the screen circuit (marked R_s in Fig. 143) is required to prevent the screen by-pass condenser from likewise by-passing the bias resistor to ground through the power supply.

In the circuits of *e*, *g*, and *h* there is a multiple feedback since, in addition to the main feedback from output to input, the omission of by-pass condensers from the cathode-biasing resistors also introduces feedback in the individual stages. In the circuit *b* the arrangement is such that the grid bias is not affected by the choice of R_2. In order to do this, C is designed to be a by-pass to ground while R_1 is an isolating resistor for a-c voltages and should be much larger than R_2. The circuit at *h* involves a bridge arrangement such that feedback can take place without introducing the possibility of direct-energy transfer between input and output circuits.

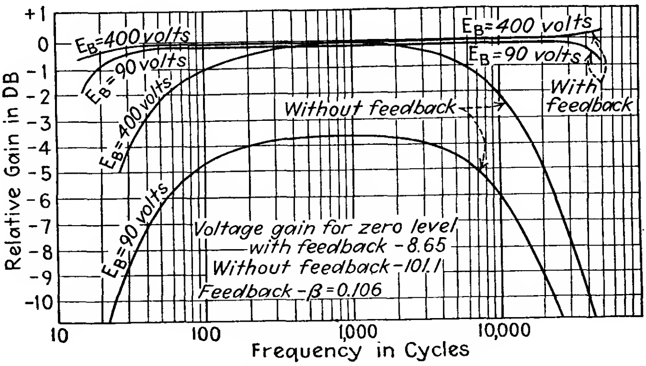


FIG. 144.—Amplification characteristic of feedback amplifier, showing effect of different amounts of feedback upon frequency distortion and constancy of amplification.

The feedback factor can be readily calculated for the cases shown in Fig. 143. Thus at *c*, *d*, and *f* the fraction β of the output voltage that is fed back to the input is $R_2/(R_1 + R_2)$, while in *a* and *e* the fraction β in the middle range of frequencies is R_2/R_{eq}, where R_{eq} is the resistance formed by grid-leak and coupling resistance in parallel.

The characteristics of a typical amplifier with negative feedback are shown in Fig. 144, which very clearly demonstrates the enormous improvement in uniformity of frequency response and in stability of gain with respect to variations in electrode voltages, which result from the proper use of negative feedback. This particular amplifier employed the circuit of Fig. 143*b* and was provided with suitable values of L_2 and C_2 to give a uniform frequency response.

The feedback principle is of practical value under many circumstances. Thus it is widely used to reduce amplitude and cross-modulation effects in power amplifiers, particularly when pentode and beam tubes are used. The improvement in frequency response and the reduction in phase shift which feedback makes possible can also be used to advantage in amplifiers intended for television and laboratory purposes. The stability of the amplification which results from the proper use of feedback is particularly

useful in laboratory measuring equipment, since with large values of $A\beta$ the amplification will stay constant with changes in electrode voltages, tube replacements, etc., to about the same extent that an ordinary direct-current voltmeter will maintain its calibration.

A modified form of negative feedback is used in some radio transmitters. This is described further in Sec. 105.

53. Behavior of Vacuum Tubes at Ultra-high Frequencies.[1]—In all the discussion that has so far been devoted to amplification it has been assumed that the time required for an electron to travel from the cathode to the plate is negligible compared with the length of time represented by a cycle, so that one could consider that the effect of changes in electrode voltages on the current were instantaneous. At very high frequencies this assumption is no longer permissible.

The result of a finite time for the transit of the electrons modifies the tube characteristics in a number of important respects. In particular, the tube constants such as the amplification factor, mutual conductance (or transconductance), and the plate resistance become vector quantities with a magnitude and a phase that vary with frequency. The mutual conductance or transconductance, for example, lags in phase because the finite transit time causes changes in the plate current to lag behind changes in grid voltage, and the absolute magnitude of the mutual conductance is also modified slightly. The amplification factor tends to decrease in magnitude and to have an increasingly large phase angle as the frequency increases, while the plate resistance likewise depends upon the frequency.

Analysis shows that the usual equivalent circuits of the amplifier tube still hold provided the amplification factor, mutual conductance, and plate resistance are given the appropriate values (including both magnitude and phase) which take into account the effect of the finite transit time.

The finite transit time also causes power to be consumed by the grid of the tube even when the grid is biased negatively and attracts no electrons. This comes about as a result of the interchange of energy between the signal voltage acting on the grid and the electrons traveling to the plate, and it is of such a character as to absorb power from the signal voltage. The amount of power lost in this way can be expressed in terms of an equivalent resistance shunted between the grid and cathode electrodes of the tube. Analysis shows that this equivalent input

[1] For further discussion of tube behavior at very high frequency see: F. B. Llewellyn, Vacuum Tube Electronics at Ultra-high Frequencies, *Proc. I.R.E.*, vol. 21, p. 1532, November, 1933; note on Vacuum Tube Electronics at Ultra-high Frequencies, *Proc. I.R.E.*, vol. 23, p. 112, February, 1935; Phase Angle of Vacuum Tube Transconductance at Very High Frequencies, *Proc. I.R.E.*, vol. 22, p. 947, August, 1934.

resistance in the case of Class A amplifier tubes is given by the formula[1]

$$\left.\begin{array}{l}\text{Equivalent input resistance resulting}\\ \text{from finite transit time}\end{array}\right\} = \frac{1}{KG_m f^2 \tau^2} \qquad (122)$$

where

G_m = mutual conductance of the tube
f = frequency
τ = time required for the electron to travel from the cathode to the grid plane.

The constant K is determined by the grid and plate voltages and by the ratio of transit times from cathode to grid plane and grid plane to anode. Consideration of the relations involved in Eq. (122) shows that the input resistance is directly proportional to the square root of the electrode voltages and inversely proportional to the square of the linear dimension of the tube.

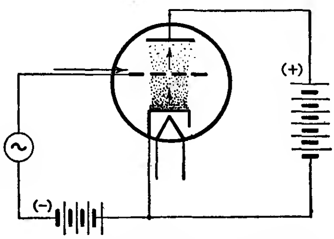


FIG. 145.—Diagram illustrating effect of transit time on grid losses. At the instant shown, the grid potential is increasing and there is a consequent disproportionate number of electrons between grid and cathode (shown by density of dots) so that a current flows into the grid.

The mechanism whereby there is loss in the grid circuit even when the grid is biased sufficiently negative so that it attracts no electrons can be understood by the following reasoning:[2] Consider a triode with a negative grid upon which is superimposed a small alternating voltage, as shown in Fig. 145. During the portion of the cycle when the instantaneous grid potential is becoming less negative, the number of electrons flowing to the plate is increasing, but the electron density is proportionately greater on the cathode side of the grid than on the anode side since the finite transit time delays the arrival of the electrons at the plate. The grid under such conditions draws current as a result of the electrostatic charges which the excess of approaching over receding electrons induces on the grid structure. Somewhat later in the cycle, when the instantaneous grid voltage is decreasing, the opposite situation exists, as there is

[1] For a theoretical derivation of Eq. (122), together with a formula for calculating the constant K, see D. O. North, Analysis of the Effects of Space Charge on Grid Impedance, *Proc. I.R.E.*, vol. 24, p. 108, January, 1936.

For an experimental verification of Eq. (122), see W. R. Ferris, Input Resistance of Vacuum Tubes as Ultra-high Frequency Amplifiers, *Proc. I.R.E.*, vol. 24, p. 82, January, 1936. In a discussion of this latter paper, J. G. Chaffee gives experimental data on vacuum-tube amplifiers biased to cut-off, which indicates that under these conditions the input resistance is inversely proportional to the first power of the frequency rather than the second power.

[2] This is due to W. R. Ferris, *loc. cit.*

now a disproportionately small number of electrons approaching the grid from the cathode in relation to the number receding from the grid toward the anode. This comes about because of the finite transit time which prevents the electrons from reaching the plate instantly. The net result of this excess of receding electrons is to cause current to flow out of the grid. The magnitude of this current flowing in and out of the grid as a result of the finite transit time is proportional to the number of electrons involved (*i.e.*, to the mutual conductance), to the signal voltage applied to the grid, to the transit time, and to the frequency.

From the above discussion it might be thought that the effect of the finite transit time would be merely to introduce a reactive grid current. This is only partially true, however, because the phase of this alternating current flowing to the grid is not quite in quadrature with the signal voltage applied to the grid. Consider the instant when the grid potential is at its least negative value (maximum positive value of applied voltage). Because of the finite transit time there is still a disproportionately large number of electrons on the cathode side of the grid as compared with those on the anode side. There is hence still current flowing into the grid at the crest of the cycle, thereby giving the grid current a component that represents power loss. The phase displacement between the maximum of signal voltage and the moment of zero grid current is proportional to the product of frequency and transit time, *i.e.*, to the number of electrical degrees represented by the time it takes an electron to travel from the cathode to the grid plane. Since this number of electrical degrees is small, and the phase angle of the grid current is proportional to $f\tau$, and the grid resistance is inversely proportional to $G_m f^2 \tau^2$.

The grid loss occasioned by finite transit time is negligible with standard tubes at broadcast frequencies, but becomes increasingly important at higher frequencies. Thus measurements upon a typical Type 57 pentode show that at 1000 kc the input resistance is approximately 21 megohms, but, since the input resistance is inversely proportional to the square of the frequency, at 30 mc it becomes approximately 23,000 ohms and at 100 mc is only 2100 ohms! As a result, the maximum possible amplification obtainable from a 57 tube at 30 mc is approximately 15, and the amplification falls to less than unity at about 100 mc.

Small dimensions are a great help in raising the input resistance at high frequencies. Thus the "acorn" tubes have dimensions approximately one-fourth to one-fifth those of standard tubes, and accordingly have input resistances about 20 times as high at the same frequency, or have an input resistance at least as high as that of a standard tube at a frequency four to five times as great. Increasing the electrode potentials also helps somewhat but dissipation in the tube limits the maximum permissible voltages.

54. Miscellaneous. *Equalization of Amplifier Gain.*—Where several stages of amplification are employed in cascade, the way in which the amplification of the individual stages varies with frequency is not particularly important as long as the individual stages combine to give the desired over-all result. Thus an excessive falling off of the high-frequency response of one stage can be compensated for by introducing a corresponding high-frequency peak in another stage of amplification, as illustrated in Fig. 146. This possibility of compensating for the deficiencies in over-all frequency response of a number of stages by modifying the characteristics of only one of the stages simplifies greatly the problem of obtaining a desired result. In particular, it avoids the cost and trouble involved in attempting to make each individual stage perfect within itself.

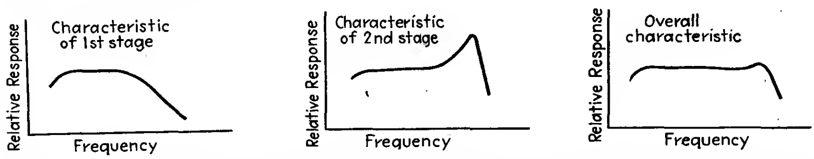


FIG. 146.—Curves illustrating how the deficiencies in the frequency response of one stage of an amplifier may be compensated for by the characteristic of another stage to give a uniform over-all response.

Use of Decibels to Express Relative Amplification.—The variation of amplification with frequency in audio-frequency amplifiers is often expressed in decibels referred to some arbitrary level taken as zero decibels. The significance of such curves can be understood by considering what a decibel means. The decibel is a unit for expressing a power ratio and is given by the relation

$$\text{Decibels} = \text{db} = 10 \log_{10} \frac{P_2}{P_1} \tag{123}$$

The decibel has no other significance; if it is to be used in expressing relative amplification, it therefore signifies power output as a function of frequency with respect to some arbitrary power. Thus, if the output voltage varies with frequency as shown in Fig. 147*a*, one might replot this curve in decibels by assuming some arbitrary power as the standard. This might, for instance, be the power output obtained at 400 cycles. The power output at any other frequency is then proportional to $(E/E_{400})^2$ where E is the voltage output at the frequency in question and E_{400} is the output voltage at 400 cycles. Since the power output under these conditions is proportional to the square of the voltage, one can rewrite Eq. (123) as follows for this particular case:

$$\text{db} = 10 \log_{10} \frac{P_2}{P_1} = 10 \log_{10} \left(\frac{E}{E_{400}}\right)^2 = 20 \log_{10} \left(\frac{E}{E_{400}}\right) \tag{124}$$

It is now possible to plot a curve giving amplification in terms of decibels as is done at Fig. 147*b*. The significance of the decibel curve can be seen by considering a specific case. Thus the fact that the amplification in Fig. 147*b* is 5 db lower at 45 cycles than at 400 cycles means that the output power at 45 cycles is 0.316 times the power at 400 cycles.

From the foregoing it is seen that anything which increases or decreases the amplification can have its effect expressed in terms of decibels. Thus, if one introduces an extra stage of amplification which increases the output voltage 20 times, then the gain in output power is $(20)^2$ or

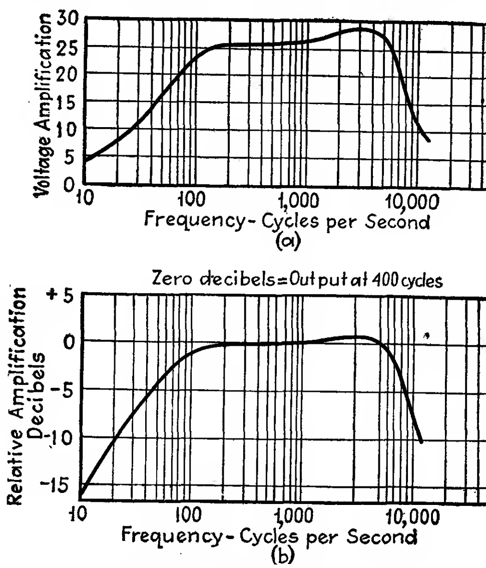


FIG. 147.—Illustration of how relative gain can be expressed in decibels.

400 times, and this power ratio, when substituted in Eq. (123), is seen to represent a power gain of 26 db.

It will be noted that the decibel is fundamentally a power unit. It cannot be used to express voltage ratios except insofar as these voltages are related to power ratios. If two voltages are applied to identical resistances, then the resulting powers are, of course, proportional to the square of the voltages; but if the voltages are applied to different resistances, then it is necessary to take into account this fact if the decibel unit is to be employed.

Power levels in audio-frequency amplifiers are often expressed in decibels referred to a standard level of 6 milliwatts. Thus an output transformer rated for service at +36-db level is capable of handling a power output that is 36 db above 6 mw, or 24 watts. Likewise, a microphone rated at −50 db will deliver an output power 50 db less than

6 mw, or 6×10^{-8} watts, when spoken into under average conditions. If the microphone has a 200-ohm internal impedance, this power will be developed in a load resistance of 200 ohms and represents an r.m.s. voltage of $(6 \times 10^{-8} \times 200)^{1/2} = 0.0035$ volt, which can then be increased to the desired level by the use of an amplifier provided with a step-up input transformer to match the grid of the first tube to the 200-ohm microphone.

Phase Distortion.—While phase distortion is of little importance with audio-frequency amplifiers, it is interesting to note the type of phase shifts obtained with different types of coupling. The phase shifts for several representative amplifiers are shown in Fig. 148, where it is seen

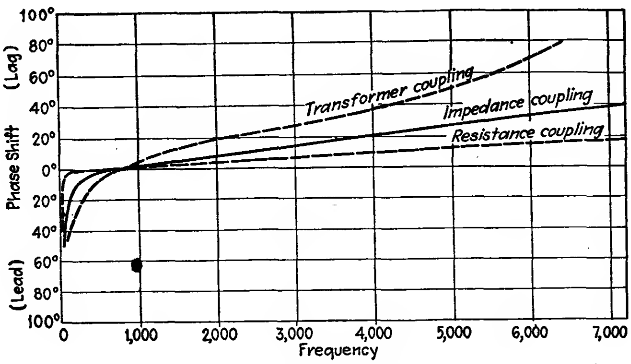


FIG. 148.—Phase shift as a function of frequency for typical amplifiers. Since absence of phase distortion requires that the phase-lag curve be a straight line passing through zero or some integral multiple of π radians at zero frequency, it is apparent that all of these curves show phase distortion.

that all the amplifiers have some phase distortion, since the curves of phase shift are not straight lines passing through zero or some integral multiple of π radians at zero frequency. The phase distortion is ordinarily least for those amplifiers having the lowest frequency distortion, and with any particular amplifier the phase distortion becomes large at frequencies where frequency distortion is pronounced.

The phase shift is related to the time required for transmission through the amplifier, or delay time, according to the formula

$$\left.\begin{array}{r}\text{Delay time}\\ \text{in seconds}\end{array}\right\} = \frac{d\beta}{d\omega} \qquad (125)$$

where β is the phase shift in radians and ω is 2π times frequency. The delay time for any frequency is hence proportional to the slope of the phase-shift curve at that particular frequency. From Fig. 148 it is apparent that with audio amplifiers the time delay tends to be constant and small over the middle-frequency range of the amplifier but is greater

at both high and low frequencies. Low frequencies give particularly large time delays because a fraction of a cycle phase shift represents much more time at a low than at a high frequency.

In tuned amplifiers the phase shift varies with frequency in a manner such as shown in Fig. 149. In the immediate vicinity of resonance the phase shift is a linear function of frequency, but begins to be appreciably non-linear when far enough off resonance for the phase shift to exceed about 30°. The time delay calculated from the phase-shift curve of a tuned amplifier in the vicinity of resonance gives the time delay that the modulation envelope suffers. This is of the same order of magnitude as the time delay commonly encountered in a single stage of audio-frequency amplification.

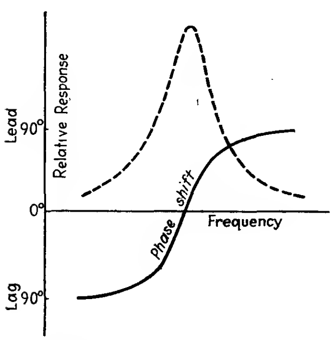


FIG. 149.—Phase shift in typical tuned amplifier. Note that the phase-shift curve is linear only in the immediate vicinity of resonance.

When several stages of amplification are in cascade, the total phase shift and time delay are obtained by adding the phase shifts and time delays, respectively, of the individual stages.

Voltage Surges in Audio-frequency Apparatus.—When the direct current flowing through the primary of an audio-frequency transformer is suddenly interrupted, there are high surge voltages developed across the transformer primary and secondary terminals which are caused by the same action that develops a high voltage across the field of an electrical machine when the field current is suddenly interrupted. This voltage may reach surprisingly large values in the case of ordinary audio-frequency transformers because, although the current interrupted is small, the inductance through which it flows is very great. The potentials obtained under ordinary operating conditions are in the neighborhood of several thousand volts and are often sufficient to produce short sparks in a gap across the transformer secondary. Surges of this sort also occur whenever a lighted tube is removed from its socket, or when stages are switched in and out of the amplifier. The voltage developed by these surges frequently causes the breakdown of insulation in the transformer, with resultant microphonic effects that appear as noises in the amplifier output.[1]

[1] For an extended discussion of surges in audio-frequency amplifiers, see E. H. Fisher, Voltage Surges in Audio-frequency Apparatus, *Proc. I.R.E.*, vol. 17, p. 841, May, 1929. It is shown in this paper that, if i_p is the transformer primary current that is suddenly broken, L_p the inductance of the transformer primary, and C_s the equivalent capacity that can be considered as acting across the secondary terminals,

Direct-current Amplifiers.—The amplification of direct-current voltages (or voltages of extremely low frequency) requires a resistance coupling between the output of one amplifier stage and the input of the following amplifier tube without the use of grid-blocking condensers. One possible arrangement of this type is shown at Fig. 150*a*, in which the voltage drop across the coupling resistance is impressed directly on the grid of the succeeding tube, which is provided with a grid-bias potential sufficient to balance out the voltage drop across the coupling resistance

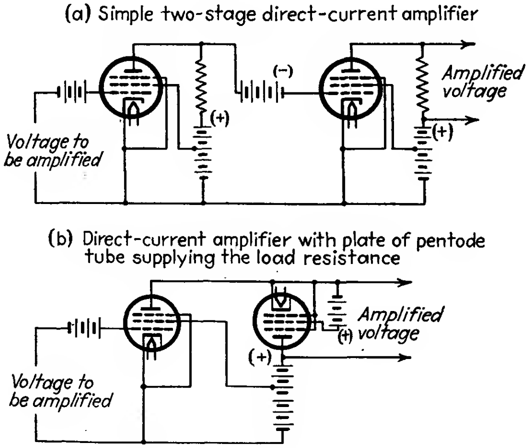


FIG. 150.—Direct-current amplifier circuits.

resulting from the normal plate current that passes through this resistance and to maintain the grid slightly negative.

The voltage gain of a direct-current amplifier can be made very large by using the plate circuit of a pentode tube as the coupling resistance, as shown in Fig. 150*b*. Such an arrangement permits a reasonable d-c plate current for the amplifier tube without an excessive loss of voltage in the load, and at the same time provides a load resistance of the order of megohms. A voltage gain of over 2000 per stage can be readily obtained in this way for direct currents and low-frequency alternating currents.[1]

---

then the maximum voltage that will be developed across the secondary terminals is given with good accuracy by the relation

$$\text{Voltage across secondary} = i_p\sqrt{\frac{L_p}{C_s}}$$

[1] For further information on such amplifiers, see J. M. Horton, The Use of a Vacuum Tube as a Plate Feed Impedance, *Jour. Franklin Inst.*, vol. 216, December, 1933.

Direct-current amplifiers of all types tend to be troubled by slow drifts in the d-c plate current, since any change in plate current caused by such effects as changes in electrode potentials cannot be distinguished from the effects produced by the direct-current voltages being amplified. A number of balancing arrangements have been devised which overcome this trouble, and these are described in detail in the literature.[1]

Reflex Amplifiers.—A reflex amplifier is essentially an ordinary audio-frequency amplifier to which tuned circuits have been added in such a way that the tube can function simultaneously as a tuned radio-frequency amplifier and as an audio-frequency amplifier. The great difference in frequency between the audio- and radio-frequency currents is utilized to obtain satisfactory separation between the two in the output. Reflex amplifiers make possible a reduction in the number of tubes, but have the disadvantage that curvature of the tube characteristics causes the amplification of the radio frequency to be affected by the audio-frequency amplitude, and vice versa, giving rise to cross-talk and distortion. For these reasons reflex amplifiers are used only under special circumstances.

55. Amplitude Distortion and Cross Modulation in Amplifiers.—In an ideal amplifier the output wave has exactly the same wave shape as the input signal. Actual amplifiers closely approach this ideal when the signal voltage being amplified is very small, but tend to depart from it and thereby introduce distortion when the signal voltage is large. As a result of this distortion, the output contains harmonics and other spurious frequencies, there may be a lack of proportionality between the amplitudes of the input voltage and the undistorted part of the output, and cross-modulation may occur between different frequency components of the signal.

When the signal is a simple sine wave, the distortion in the output wave and the exact output voltage and power can be determined by the use of the dynamic characteristic of the amplifier, as discussed in Sec. 56. This is the method commonly used to determine the proper operating conditions for minimum distortion and maximum output for voltage amplifiers that must develop large output voltages. While quite satisfactory for this purpose, the dynamic characteristic method is at best only approximate, and can be applied only to resistance loads with a sine-wave signal. A more comprehensive picture of the entire problem of distortion and cross-modulation is obtained by expressing the tube characteristic in terms of a power series, as explained in Sec. 37, and using this to obtain the relation between the output and input waves.

[1] A discussion of the characteristics and relative merits of a number of such balancing arrangements is given by D. B. Penick, Direct Current Amplifier Circuits for Use with the Electrometer Tube, *Rev. Scientific Instruments*, vol. 6, p. 115, April, 1935.

It was shown in Sec. 37 that, for electrode voltage such that the instantaneous plate current never became zero, the characteristics of a tube could be expressed in terms of the following power series

$$i_p = G_m\left(e_g + \frac{e_p}{\mu}\right) + \frac{1}{2!}\frac{\partial G_m}{\partial E_g}\left(e_g + \frac{e_p}{\mu}\right)^2 + \frac{1}{3!}\frac{\partial^2 G_m}{\partial E_g^2}\left(e_g + \frac{e_p}{\mu}\right)^3 + \cdots \quad (126)$$

where

e_g = signal voltage applied to grid

e_p = voltage acting on plate as result of voltage drop of i_p in load impedance

G_m = mutual conductance at operating point

μ = amplification factor

i_p = plate current caused by the action of e_g and e_p.

The tube constants and the derivatives appearing in Eq. (126) are evaluated at the operating point upon which the signal voltage is superimposed, and the convention with respect to sign made in deriving Eq. (126) is such that a positive current is a current flowing in the same direction as the normal d-c plate current. Equation (126) assumes either that the amplification factor μ is constant or that the load impedance is so low compared with the plate resistance that e_p/μ is negligible compared with e_g. The equation therefore holds reasonably well for triodes in the normal operating range, where the amplification factor is substantially constant, and is highly accurate for pentode amplifier tubes since pentodes normally have a very high plate resistance compared with the load resistance.

Since the voltage e_p appearing in Eq. (126) represents the voltage arising from the current i_p in flowing through the load impedance Z, one can substitute the relation $e_p = -i_pZ$ in Eq. (126). When this is done and the series is inverted to give an explicit expression for i_p, the result is[1]

[1] The transformation from Eq. (126) to Eq. (127) is carried out as follows: The substitution of $e_p = -Zi_p$ into Eq. (126) gives

$$i_p = G_m\left(e_g - \frac{i_pZ}{\mu}\right) + \frac{1}{2!}\frac{\partial G_m}{\partial E_g}\left(e_g - \frac{i_pZ}{\mu}\right)^2 + \cdots \quad (128)$$

This gives i_p as an implicit function of i_p, whereas what is wanted is an explicit solution for i_p of the form

$$i_p = a_1e_g + a_2e_g^2 + a_3e_g^3 + \cdots \quad (129)$$

The inversion from Eq. (128) to Eq. (129) is accomplished by substituting the value of i_p in the form of Eq. (129) for i_p in Eq. (128), and then equating the coefficients of like powers of e_g on both sides of the result to determine the values that the a's must

$$i_p = \frac{\mu e_g}{R_p + Z_1} + \frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}\frac{e_1^2}{R_p + Z_2} + \frac{\left(\frac{1}{3!\mu^2 G_m}\frac{\partial^2 G_m}{\partial E_g^2}e_1^3 - \frac{1}{\mu G_m}\frac{\partial G_m}{\partial E_g}e_1 e_2\right)}{R_p + Z_3} + \cdots \qquad (127)$$

where

R_p = plate resistance at the operating point

Z = load impedance. The subscripts 1, 2, and 3 denote the impedance offered to the first-, second-, and third-order components of the plate current. When any one order plate current contains components of several frequencies, the appropriate value of Z must be used for each component

$e_1 = \mu e_g\left(\frac{R_p}{R_p + Z_1}\right)$ = voltage drop produced in plate resistance by first-order component of current

$e_2 = \left(\frac{Z_2}{R_p + Z_2}\right)\frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}e_1^2$ = voltage drop produced across the load impedance Z_2 by the second-order component of current.

The remaining notation is as used previously. In the case of pentode tubes Z/R_p is usually negligible, so that to the extent this is true Eq. (127) reduces to

$$i_p \text{ (for pentodes)} = G_m e_g + \frac{1}{2!}\frac{\partial G_m}{\partial E_g}e_g^2 + \frac{1}{3!}\frac{\partial^2 G_m}{\partial E_g^2}e_g^3 + \cdots \qquad (131)$$

Examination of Eq. (127), or of Eqs. (129) and (130), shows that the actual current that flows in the plate circuit as a result of the signal voltage e_g consists of a series of components, proportional, respectively, to the first, second, third, etc., powers of the signal voltage involved. The total plate current is the sum of these various components plus the

---

have in Eq. (129) to satisfy Eq. (128). Doing this gives

$$\begin{aligned} a_1 &= \frac{\mu}{R_p + Z_1} \\ a_2 &= \frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}\frac{\left(\frac{\mu R_p}{R_p + Z_1}\right)^2}{R_p + Z_2} \\ a_3 &= \frac{\frac{1}{3!\mu^2 G_m}\frac{\partial^2 G_m}{\partial E_g^2}\left(\frac{\mu R_p}{R_p + Z_1}\right)^3 - \frac{1}{\mu G_m}\frac{\partial G_m}{\partial E_g}\left(\frac{\mu R_p}{R_p + Z_1}\right)(a_2 Z_2)}{R_p + Z_3} \end{aligned} \qquad (130)$$

Equation (127) now follows by substituting these values of a into Eq. (129). In Eq. (130) the impedances Z_1, Z_2, and Z_3 represent the impedances offered by the load to the components a_1e_g, $a_2e_g^2$, and $a_3e_g^3$, respectively, of the plate current.

steady d-c plate current present at the operating point when no signal is applied to the grid.

Study of Eq. (127) shows that the various components of the plate current can be considered as being produced by a series of equivalent voltages acting in a circuit consisting of the plate resistance of the tube in series with the load impedance. If the positive direction is taken as from cathode toward plate [the opposite of the positive direction in Eq. (127)], the equivalent voltage represented in the first term on the right-hand side of Eq. (127) is

$$\left.\begin{matrix}\text{Equivalent voltage producing}\\ \text{first-order component of plate}\\ \text{current}\end{matrix}\right\} = -\mu e_g \qquad (132)$$

This voltage is that used in Sec. 43 in setting up the equivalent circuit of the vacuum-tube amplifier, which now by a rigorous analysis is shown to be the first-order approximation that results when the characteristics of the tube are expressed in terms of a power series. The equivalent voltage represented in the second term on the right-hand side of Eq. (127), assuming the positive direction is from cathode toward plate, is

$$\left.\begin{matrix}\text{Equivalent voltage producing}\\ \text{second-order component of}\\ \text{plate current}\end{matrix}\right\} = -\frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}\left(\mu e_g \frac{R_p}{R_p + Z_1}\right)^2 = -\frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}e_1^2 \qquad (133)$$

This leads to the equivalent circuit of Fig. 151*b*, which takes into account second-order effects. These second-order effects represent to a first approximation the error involved when all components except the first are neglected. The third term on the right-hand side of Eq. (127) represents an equivalent voltage which, when the positive direction is assumed to be from the cathode toward the plate, is given by

$$\left.\begin{matrix}\text{Equivalent voltage produc-}\\ \text{ing third-order component}\\ \text{of plate current}\end{matrix}\right\} = -\frac{1}{3!\mu^2 G_m}\frac{\partial^2 G_m}{\partial E_g^2}e_1^3 + \frac{1}{\mu G_m}\frac{\partial G_m}{\partial E_g}e_1 e_2 \qquad (134)$$

This leads to the equivalent circuit of Fig. 151*c*, which takes into account third-order effects. The third-order component of the plate current represents to a first approximation the error involved when all components except the first and second are neglected. The analysis could be continued to higher order components with similar results, but this will not be done since terms higher than the third order are not particularly important.

To summarize the above, it is seen that the behavior of the amplifier tube, including the distortion effects produced when large signal voltages

are applied, can be determined by postulating a series of generators acting in a circuit consisting of the plate resistance of the tube in series with the load impedance as shown at Fig. 151*d*. There is one such generator for each order effect, with the value of the generator voltage given by the expressions indicated in Fig. 151. The only assumptions involved in the equivalent circuits of Fig. 151 and the analysis given above are that the plate current does not reach zero or the full saturation value, and that the amplification factor is constant (or that the load impedance is small compared with the plate resistance).

Application of Power-series Method to the Analysis of Distortion in Amplifiers.—The nature of the distortions that occur in amplifiers can be

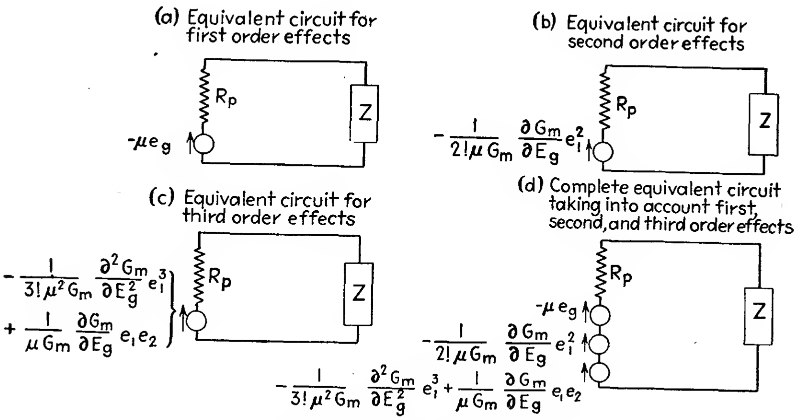


FIG. 151.—Equivalent circuits taking into account first-, second-, and third-order components of the plate current.

determined by considering the equivalent voltages that can be considered as acting in the plate circuit under various conditions.

The first-order component of the plate current needs no discussion since it represents the desired part of the output wave which is free of amplitude distortion and which has been used throughout this and the preceding chapters.

The second-order equivalent voltage is proportional to e_1^2 and can be investigated by writing down e_1 as a function of time and squaring the result. When the signal voltage e_g is a sine wave, then e_1 is likewise a sine wave of the form $e_1 = E_a \sin \omega t$. The equivalent second-order voltage is then, by Eq. (133),

$$\left.\begin{array}{l}\text{Equivalent second-}\\\text{order voltage}\end{array}\right\} = -\frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}E_a^2 \sin^2 \omega t$$

$$= -\frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}E_a^2\left(\frac{1 - \cos 2\omega t}{2}\right) \qquad (135a)$$

The right-hand side of Eq. (135a) contains two parts, one a direct-current term and the other a second harmonic of the signal frequency. These two are of equal amplitude and are proportional to the square of the signal voltage. When the signal consists of the sum of two sine waves of different frequencies, then e_1 is of the form $e_1 = E_a \sin \omega t + E_b \sin vt$, and the equivalent second-order voltage becomes

$$\left.\begin{array}{l}\text{Equivalent second-}\\ \text{order voltage}\end{array}\right\} = -\frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}(E_a \sin \omega t + E_b \sin vt)^2$$

$$= -\frac{1}{2!\mu G_m}\frac{\partial G_m}{\partial E_g}\left(\frac{E_a{}^2 + E_b{}^2}{2} - \frac{E_a{}^2 \cos 2\omega t + E_b{}^2 \cos 2vt}{2} + E_aE_b \cos(\omega + v)t + E_aE_b \cos(\omega - v)t\right) \quad (135b)$$

The right-hand side of Eq. (135b) now contains a direct-current term proportional to the square of the effective value of the signal, second-harmonic terms of each signal frequency proportional to the square of the parts of the signal involved, one term proportional to the product of the amplitudes of the two components of the signal and having a frequency that is the difference of the two frequencies, and a companion term of the same amplitude but having a frequency that is the sum of the frequencies of the two parts of the signal.

The third-order equivalent voltage contains two components, one proportional to e_1^3 and the other proportional to e_1e_2 and hence produced by the interaction of the first- and second-order currents in the plate circuit. When the load impedance is small compared with the plate resistance, as in the case of pentodes, this second part of the third-order equivalent voltage is negligible. When the signal voltage is a sine wave, then e_1 is likewise a sine wave of the form $e_1 = E_a \sin \omega t$. The first part of the third-order equivalent voltage is then, by Eq. (134),

$$\left.\begin{array}{l}\text{First part of equiva-}\\ \text{lent third-order volt-}\\ \text{age}\end{array}\right\} = -\frac{1}{3!\mu^2 G_m}\frac{\partial^2 G_m}{\partial E_g{}^2}E_a{}^3 \sin^3 \omega t$$

$$= -\frac{1}{3!\mu^2 G_m}\frac{\partial^2 G_m}{\partial E_g{}^2}E_a{}^3\left(\frac{3}{4}\sin \omega t - \frac{1}{4}\sin 3\omega t\right) \quad (136a)$$

The right-hand side of Eq. (136a) contains two components, both proportional to the cube of the signal voltage and having frequencies equal to and three times the signal frequency, respectively. The first of these components combines with the undistorted part of the output arising from the first-order equivalent voltage, but, since this additional contribution is proportional to the cube of the signal instead of the first power, the result is a loss in proportionality between the input and output.

When the signal voltage consists of the sum of two sine waves, then e_1 is of the form $e_1 = E_a \sin \omega t + E_b \sin vt$, and the first part of the equivalent third-order voltage is, by Eq. (134),

$$\left.\begin{array}{l}\text{First part of equiva-}\\ \text{lent third-order vol-}\\ \text{tage}\end{array}\right\} = -\frac{1}{3!\mu^2 G_m}\frac{\partial^2 G_m}{\partial E_g^2}(E_a^3 \sin^3 \omega t + 3E_a^2 E_b \sin^2 \omega t \sin vt + 3E_a E_b^2 \sin \omega t \sin^2 vt + E_b^3 \sin^3 vt)$$

$$= -\frac{1}{3!\mu^2 G_m}\frac{\partial^2 G_m}{\partial E_g^2}\left\{\left(\frac{3}{4}E_a^3 + \frac{3E_aE_b^2}{2}\right)\sin \omega t + \left(\frac{3}{4}E_b^3 + \frac{3E_a^2E_b}{2}\right)\sin vt - \frac{E_a^3 \sin 3\omega t + E_b^3 \sin 3vt}{4} - \frac{3}{4}E_a^2E_b[\sin(2\omega + v)t + \sin(-2\omega + v)t] - \frac{3}{4}E_aE_b^2[\sin(\omega + 2v)t + \sin(\omega - 2v)t]\right\} \quad (136b)$$

The right-hand side of Eq. (136b) contains third-harmonic terms of each of the signal frequencies, and also third-order combination frequencies such as $\omega \pm 2v$ and $2\omega \pm v$. It also contains terms having the same frequency as the signal, which combine with the undistorted part of the output arising from the first-order equivalent voltage. These terms, however, prevent the part of the output having the same frequency as the signal from being proportional to the signal voltage. Furthermore, output obtained for any one component frequency of the signal is also dependent to some extent upon the amplitude of the second frequency component of the signal. This action is called *cross-modulation* and is discussed below.

Examination of Eq. (134) shows that the second part of the equivalent third-order voltage (the part proportional to e_1e_2) contains the same frequency components as the first part, and hence merely alters the magnitude of the third-order distortion without affecting its character.

The above discussion can be readily extended to cover the case in which the signal voltage is a modulated wave by assuming that e_1 is of the form $e_1 = E_a \sin \omega t = E_0(1 + m \sin \gamma t) \sin \omega t$, in which E_a now varies according to the modulation envelope. The first-order component of the plate current then produces an output wave that is an exact reproduction of the signal, just as before. The second-order action gives rise to a direct-current term, and likewise a second-harmonic carrier frequency having a modulation that is a distorted replica of the modulation of the original signal. The third-order action gives rise to a third-harmonic carrier also having distorted modulation, and along with this is a component having the same carrier frequency as the signal,

but with a distorted modulation. This latter component is the only higher order effect to which the resonant circuit of a tuned radio-frequency amplifier responds, and *it combines with the first-order output to give an output voltage having a distorted modulation.*

When the signal consists of two modulated waves, the second-order effects give rise to direct-current, second-harmonic, and sum and difference frequency terms in the output, but none of these are important because an amplifier tuned to the signal frequency will not respond to them. The third-order action is quite important, however, because, in addition to third harmonics and third-order combination frequencies to which the tuned output circuit will not respond, there are components having the same frequency as the signal but with distorted modulations, giving rise to distortion in the modulation envelopes of the output waves. There is also cross-modulation between the two parts of the signal, as a result of which the modulation of one carrier frequency gets transferred to some extent to the other carrier frequency. This comes about as a result of the fact that the output on one signal component depends upon the amplitude of the other signal component. Hence, if the latter is a modulated wave, the output of the first component will vary with the modulation of the second wave. The ratio of the degree of modulation produced on the first wave by the modulation on the second wave to the modulation on the second carrier for small degrees of modulation of the latter, is termed the *cross-talk factor;* for amplifiers employing pentode tubes it is[1]

$$\text{Cross-talk factor} = \frac{E_2^2}{2G_m}\frac{\partial^2 G_m}{\partial E_g^2} \tag{137}$$

where E_2 is the crest amplitude of the carrier applied to the grid, and whose modulation is being transferred to the other carrier frequency.

[1] Equation (137) can be readily derived from Eq. (136*b*), neglecting the second part of the third-order equivalent voltage because it is negligible in pentodes. It is noted in Eq. (136*b*) that the cross-modulation of the E_a component on the E_b component arises from the term $\frac{1}{3!\mu^2 G_m}\frac{\partial^2 G_m}{\partial E_g^2}\frac{3E_a^2E_b}{2}\sin vt$. Writing $E_a = \mu E_2(1 + m\sin\gamma t)$, where E_2 is the carrier of the E_a voltage component reduced to the grid (note that E_a by definition is referred to the plate circuit), then, if m is small, one has to a good approximation

$$E_a^2 = \mu^2E_2^2(1 + 2m\sin\gamma t)$$

Hence the above third-order term becomes $\frac{1}{3!G_m}\frac{\partial^2 G_m}{\partial E_g^2}\frac{3}{2}E_2^2E_b(1 + 2m\sin\gamma t)\sin vt$. This causes the amplitude of the other (or E_b) carrier to vary by an amount $\frac{1}{2G_m}\frac{\partial^2 G_m}{\partial E_g^2}E_bE_2^2m\sin\gamma t$. This divided by E_b gives the degree of modulation produced on the E_b carrier, and divided again by m gives the cross-talk factor.

Summary of Effects Produced by Different Order Components of the Plate Current.—The first-order component of the plate current represents the current produced in the equivalent amplifier circuit described in Sec. 43 and used in analyzing amplifier behavior. This component represents the undistorted part of the output.

The second-order component of the plate current includes a direct-current component, second harmonics of the signal frequencies, and sum and difference frequencies formed by combinations of signal components of different frequencies. The harmonic and sum and difference frequencies arising from second-order action are the principal sources of amplitude distortion in Class A audio amplifiers when operated to give only moderate distortion. Second-order effects are not important with tuned amplifiers because the tuned load circuit does not respond to any of the second-order products. The second-order effects are proportional to the curvature of the I_p-E_g characteristic curve of the tube at the operating point; thus they can be minimized by operating where the characteristic curves are most nearly linear. The second-order effects can also be reduced by making the load impedance high compared with the plate resistance. This condition can be readily realized in the case of triodes, and makes possible a reduction of the second-order distortion to a much lower value in triodes than is possible in pentodes.

The third-order component of the plate current includes third harmonics and third-order combination frequencies, gives rise to cross-modulation, and produces a lack of proportionality between output and input. The last two effects are the most important, particularly as they represent types of behavior not produced by lower order effects. Most of the third-order action usually results from the third derivative of the tube's characteristic curve (*i.e.*, the rate of change of curvature). Hence cross-modulation and lack of proportionality can be reduced by restricting the operating range to regions where the curvature of the tube characteristic is changing only slowly. This means, in general, avoiding the region near cut-off, or using tubes with gradual cut-off (variable-mu type).[1] Third-order effects can also be reduced by making the load impedance to the first-order currents high compared with the plate resistance. This condition can be readily realized with triodes

[1] The variable-mu tube was developed to provide a radio-frequency pentode in which the amplification could be controlled by varying the grid bias without producing excessive cross-talk and lack of proportionality at low values of mutual conductance. This is accomplished by the gradual cutoff, which makes the rate of change of curvature (the third derivative) much smaller at low values of plate current than is the case with tubes having a sharp cut-off. See Stuart Ballantine and H. A. Snow, Reduction of Distortion and Cross-talk in Radio Receivers by Means of Variable-mu Tetrodes, *Proc. I.R.E.*, vol. 18, p. 2102, December, 1930.

and means that triode tubes tend to give less trouble from cross-modulation and lack of proportionality than do pentodes.

Fourth and higher even-order effects are similar in character to the second-order effects, giving rise to even harmonics, even-order combination frequencies, and direct-current terms. In like manner the fifth and higher odd-order effects act similarly to the third order, giving rise to odd harmonics, odd-order combination frequencies, cross-talk, and lack of proportionality between input and output. These higher order terms are, however, usually much smaller in magnitude than the second- and third-order terms; thus they can be neglected except under special circumstances.

Experimental Evaluation of Second- and Third-order Coefficients.—In order to make quantitative calculations based upon the above analysis, it is necessary to evaluate $\partial G_m/\partial E_g$ and $\partial^2 G_m/\partial E_g^2$ as well as the plate impedance, mutual conductance, etc., at the operating point. The value of $\partial G_m/\partial E_g$ can be readily determined experimentally by applying to the grid of the tube a sine-wave voltage of known amplitude and by noting the change of d-c plate current that results when the load impedance in the plate circuit is negligibly small. According to Eqs. (127), (131), and (135*a*) one then has

$$\frac{\partial G_m}{\partial E_g} = \frac{4\Delta I}{E^2} \tag{138}$$

where ΔI is the change in d-c plate current and E is the crest value of the signal voltage applied to the grid.

The value of $\partial^2 G_m/\partial E_g^2$ can be determined in a similar manner by applying to the grid a sine wave of known amplitude and then measuring the third-harmonic component of the plate current when the load impedance is very small. According to Eqs. (127), (131), and (136*a*) one then has

$$\frac{\partial^2 G_m}{\partial E_g^2} = \frac{24 I_3}{E^3} \tag{139}$$

where I_3 is the crest value of the third-harmonic component of the plate current and E is the crest value of the signal voltage applied to the grid.

Problems

1. Sketch the circuit diagram of a volume-control arrangement suitable for use with impedance-coupled amplifier.

2. *a.* What would be the objections to varying the amplification in a resistance-coupled amplifier of Fig. 93 by varying the grid-leak resistance?

b. What would be the disadvantage of controlling the amplification of a transformer-coupled amplifier by varying a resistance shunted across the secondary?

3. *a.* Why is it that the amplification of a stage of radio-frequency amplification as illustrated in Fig. 120*a* cannot be varied satisfactorily by connecting the grid of the output tube to a sliding contact operating on the grid-leak resistance?

b. Discuss the problem of controlling the gain of a resistance-coupled pentode amplifier by varying the control grid bias.

4. Sketch a detailed circuit diagram for carrying out the operations indicated in Fig. 129.

5. Explain why a transformer in a shield of permalloy or similar alloy reduces the hum pick-up as compared with a cast-iron shield.

6. *a.* Calculate the total noise voltage developed across a 1-megohm resistor at 20°C. in the frequency band 40 to 12,000 cycles.

b. A condenser of 0.0001 mf is shunted across the resistor of (*a*). Calculate the total noise voltage under this condition for the same frequency band.

7. What is the *effective* value of the noise voltage in the output of an amplifier having an over-all voltage gain of 100,000 times, if the input resistance across the first grid to filament is 100,000 ohms and the amplifier gives substantially constant gain over the band 40 to 8000 cycles and very little gain outside this range?

8. In a triode resistance-coupled amplifier:

$$\begin{aligned} \mu &= 100 \\ R_p &= 125{,}000 \text{ ohms} \\ R_C &= 250{,}000 \text{ ohms} \\ R_{GL} &= 500{,}000 \text{ ohms} \\ C_C &= 0.01 \text{ mf} \\ C_s &= 60\ \mu\mu\text{f} \\ C_{gp} &= 1.7\ \mu\mu\text{f} \\ C_{gf} &= 1.7\ \mu\mu\text{f} \\ C_{pf} &= 3.8\ \mu\mu\text{f} \end{aligned}$$

Calculate the input capacity and input resistance as a function of frequency for the frequency range 1000 to 30,000 cycles.

9. Check the results of Fig. 131 by recalculating from the circuit data given in the figure.

10. Derive equations analogous to Eqs. (113) and (114) but for the case of over-neutralization, using the circuit of Fig. 134 and assuming perfect coupling between the neutralizing coil and the plate circuit inductance with a unity ratio.

11. In a two-stage resistance-coupled amplifier the second stage is identical with the amplifier of Prob. 8, while the first stage is the same except that the effective shunting capacity in its plate circuit is 10 $\mu\mu$f plus the input capacity of the second stage and that the grid-leak resistance of the second stage is shunted by the input resistance of the second stage. With the aid of the results of Prob. 8, calculate and plot the curve of amplification as a function of frequency for the first stage; for comparison, also calculate and plot the curve that would be obtained if the input admittance of the second stage were zero.

12. In a particular amplifier a signal voltage of 0.01 volt applied to the grid of the first tube causes a signal current of 5 ma to flow in the plate circuit of the output tube. The first and last tubes receive their plate voltages from a common battery that may have an internal resistance as high as 200 ohms. The first stage is transformer coupled, with a turn ratio of 3 and a tube amplification factor of 13. Devise means to prevent the feedback from affecting the amplification by more than 10 per cent for all frequencies between 50 and 8000 cycles.

13. A three-stage resistance-coupled amplifier is designed according to the first column of Table VII, page 188. If these stages all receive their plate voltage from a common source having an internal impedance consisting of a capacity of 8 mf, design a filter system so that the common coupling will not affect the amplification by more than approximately 10 per cent down to a frequency of 30 cycles.

14. In tuned radio-frequency amplifiers operating in very high frequency bands it is found that the amplifier has a tendency to oscillate unless the by-pass condensers are designed to be non-inductive, but this effect is ordinarily absent at lower frequencies. Explain.

15. *a*. Design a single-tube feedback amplifier to have a voltage amplification of exactly 10, using the circuit of Fig. 143*c* and assuming the plate-supply potential is 250 volts.

b. Calculate and plot the amplification as a function of frequency in the absence of feedback, making reasonable assumptions as to tube constants and shunting capacities.

c. Repeat (*b*), but taking into account feedback.

d. Repeat (*b*) and (*c*) for the case where the original amplifier tube is replaced by another tube having only 70 per cent as much mutual conductance, but otherwise with identical characteristics.

Plot all curves on the same curve sheet.

16. Derive Eq. (121).

17. In a particular tube the input resistance as a result of finite transit time is 500,000 ohms at 6 mc and normal electrode voltages.

a. What would be the input resistance at 60 mc with normal electrode voltages?

b. What would be the input resistance at 60 mc if the electrode voltages were doubled?

c. What would be the input resistance in (*a*) if the tube were redesigned so that all linear dimensions were halved.

18. A two-stage audio-frequency amplifier is required to have a substantially constant gain down to 20 cycles at the low-frequency end. If one of the stages is the amplifier of Fig. 93, sketch the circuit of an amplifier that could be used in the second stage to produce the necessary equalization, and discuss the basis for adjusting this second stage.

19. *a*. Plot the amplification curve of Fig. 93 in decibels, assuming that the gain at 1000 cycles is taken as zero decibels.

b. What is the gain in decibels corresponding to zero decibels in *a*?

20. Determine and plot the time delay as a function of frequency for the amplifier of Fig. 93.

21. Work out the circuit of a reflex amplifier employing a pentode tube.

22. Analysis of the characteristic curves of a particular triode tube shows that at the operating point $\mu = 10$, $R_p = 50{,}000$, $G_m = 200$ micromhos, $\partial G_m/\partial E_g = 10^{-5}$, and $\partial^2 G_m/\partial E_g^2 = -10^{-7}$.

a. Calculate the exact currents that will flow in the plate circuit up to and including third-order effects, when an alternating signal potential of 2 volts crest value is applied to the grid of the tube and the load resistance is negligible.

b. Calculate the cross-talk factor for an interfering carrier of 2 volts.

CHAPTER VII

POWER AMPLIFIERS

56. Class A Power Amplifiers Employing Triode Tubes.—In a power amplifier the object is to develop power, in contrast with the voltage amplifier, in which the object is to obtain as much voltage as possible. In the Class A power amplifier the tube is operated so that the wave shape of current in the plate circuit of the tube reproduces as nearly as possible the wave shape of the signal voltage applied to the grid of the tube. Class A power amplifiers employing triode tubes are used in nearly all radio receivers, in small public-address systems, in telephone amplifiers, etc.

The problems involved in realizing the full possibilities of a triode tube as a power amplifier can be understood by studying the voltage and current relations existing in a tube when a large signal voltage is applied to the grid. Consider the case of an amplifier tube having characteristics such as given in Fig. 152, and operated at a grid-bias and plate voltage that place the operating point at the spot marked O. When there is no load impedance in the plate circuit, a signal voltage applied to the grid causes variations in plate current that fall along the E_g-I_p characteristic curve of the tube marked a in the figure. The effect of a load impedance is to cause the plate current to follow a line such as b, having a slope less than that of the tube characteristic a by an amount depending upon the load resistance. The path b is known as the *dynamic characteristic* because it gives the behavior of the tube under actual working conditions, *i.e.*, when an alternating voltage is applied to the grid and when there is a load impedance in the plate circuit. The dynamic characteristic does not follow the tube characteristic a because, when a load impedance is present, the voltage at the plate of the tube differs from the voltage of the plate supply by the drop in the load impedance. Thus, when the signal is on the positive half cycle, the plate current increases above the average value, causing a voltage drop in the load resistance which reduces the potential actually applied to the plate to something less than the plate voltage at the operating point, while during the negative half of the signal cycle the plate current is less than the average value and the voltage drop in the resistance is such as to make the instantaneous plate voltage somewhat more than the plate voltage at the operating point. The dynamic characteristic therefore has a slope that is less than the slope of the static characteristic, and, in fact, has a slope that corresponds

to a resistance equal to the plate resistance of the tube plus the load resistance. The dynamic characteristic is also more nearly linear than are the characteristic curves of the tube.

In order to minimize distortion, it is necessary to restrict the operating range to the portion of the dynamic characteristic that is substantially linear. The practical limiting factors are the excessive curvature of the dynamic characteristic at low plate currents and the fact that, if the grid is driven positive, the grid current that results tends to distort the input wave. Trouble of the latter sort is avoided in most Class A amplifiers by limiting the crest value of the signal voltage to a value that does not exceed the grid bias, so that the instantaneous grid potential

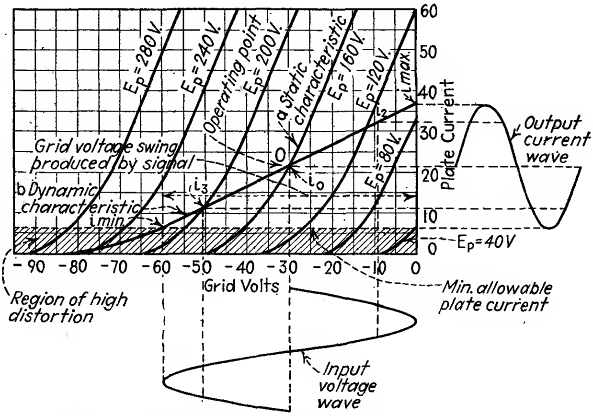


FIG. 152.—Characteristic curves of typical power amplifier tube, showing region of high distortion, together with dynamic characteristic for a load resistance that gives proper operating conditions, and waves of input voltage and output current.

is never positive. The special problems involved when the grid is intentionally driven positive are considered later. The dynamic characteristic with triode tubes having a resistance load always approximates very closely a straight line except in the region where the plate current is small. This is illustrated by Fig. 152, in which the approximate location of this region of high curvature (large distortion) is shaded. It is apparent that to avoid excessive distortion the instantaneous plate current must never fall into the shaded region.

Approximate Analysis of Relations Existing in Triode Class A Power Amplifiers When the Grid Is Not Driven Positive.—In order to obtain the maximum possible power output without making the instantaneous plate current too small during the most negative part of the applied signal and without driving the grid positive at the positive peak of the cycle, it is necessary to maintain a very careful balance between grid bias, load impedance, plate-supply voltage, and plate resistance.

While an exact mathematical treatment of the relations existing in the power tube is complicated by the curvature in the dynamic characteristic, it is possible to determine, with an accuracy that is sufficient for most practical purposes, the proper operating conditions and the undistorted power output that can be expected from a power tube, by considering that the curves showing the relation between grid voltage and plate current are straight lines for values of plate current greater than the shaded region of Fig. 152. Such a characteristic, an example of which

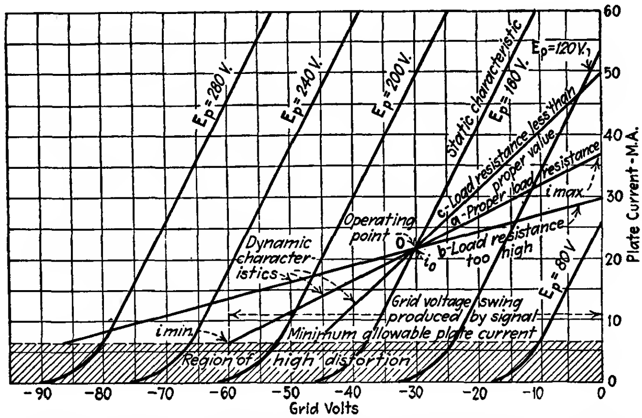


FIG. 153.—Idealized characteristic curves used to determine proper operating conditions for power amplifier, together with idealized dynamic characteristics for the proper load resistance and for load resistances greater and less than the value giving maximum power output.

is shown in Fig. 153, is subject to simple mathematical treatment and at the same time does not differ very greatly from the actual characteristic. With this approximation the proper grid bias with a resistance load is given by the equation[1]

[1] This equation, together with those that follow, can be derived in the following manner based on the idealized characteristic curves of Fig. 153. At the instant the signal is at its negative crest, the instantaneous plate current should be the minimum allowable value. In order to accomplish this, the instantaneous grid and plate potentials e_g and e_p must satisfy the relation

$$\frac{E_0}{\mu} = e_g + \frac{e_p}{\mu} \tag{140}$$

If E_s is the crest signal voltage, then, at the moment when the plate current is at its minimum, $e_g = -E_c - E_s$. At the same time e_p is the plate supply voltage minus the voltage drop in the load; as the drop is $\frac{-\mu E_s R_L}{R_p + R_L}$, then, when the signal is at its negative crest, $e_p = E_B + \frac{\mu E_s R_L}{R_p + R_L}$ Substituting these in Eq. (140) gives

$$\text{Proper negative grid bias of power tube} = \left(\frac{E_B - E_0}{\mu}\right)\left(\frac{R_L + R_p}{R_L + 2R_p}\right) \tag{144}$$

where

E_B = plate voltage at operating point

E_0 = plate voltage required to develop the minimum allowable plate current, assuming the grid-bias voltage to be zero

R_p = plate resistance of tube at operating point

R_L = load resistance

μ = amplification factor of tube.

This value of grid bias locates the operating point so that a sine-wave signal voltage reaches the limits of zero instantaneous grid voltage on the positive half cycle and minimum allowable plate current on the negative half cycle simultaneously, and so places the operating point where the power output is the maximum that can be obtained. The proper value of grid bias becomes more negative as the load resistance and the plate-supply voltage become greater.

$$\frac{E_0}{\mu} = -(E_c + E_s) + \frac{E_B}{\mu} + E_s\left(\frac{R_L}{R_p + R_L}\right)$$

or (141)

$$E_c + E_s\left(\frac{R_p}{R_L + R_p}\right) = \frac{E_B - E_0}{\mu}$$

When $E_s = E_c$, this reduces to Eq. (144).

The power output developed in the load is equal to half the square of the crest voltage developed across the load divided by the load resistance, and, as the crest voltage across the load is $\mu E_s R_L/(R_p + R_L)$, then

$$\text{Power output} = \frac{(\mu E_s)^2 R_L}{2(R_p + R_L)^2} \tag{142a}$$

This reduces to Eq. (146*a*) when $E_s = E_c$ and the value of E_c given by Eq. (144) is substituted for E_s.

The plate efficiency is the ratio of output power as given by Eq. (142*a*) to the direct-current power supplied to the tube by the source of plate voltage. This input power is $E_B I_0$, where I_0 is the d-c current at the operating point. If the ratio of minimum plate current reached during the cycle to the plate current at the operating point is denoted by Φ (*i.e.*, $\Phi = I_{\min}/I_0$), then

$$I_0(1 - \Phi) = \frac{\mu E_s}{R_p + R_L}$$

The input power then becomes

$$E_B I_0 = \frac{\mu E_s E_B}{(1 - \Phi)(R_p + R_L)}$$

The ratio of output to input is then

$$\text{Efficiency} = \frac{\mu E_s R_L (1 - \Phi)}{2E_B(R_p + R_L)} \tag{143a}$$

When $E_s = E_c$ and the value of E_c from Eq. (144) is used, the result is Eq. (149*a*).

Alternative expressions for power output and efficiency are obtained by noting

Equation (144) can be rewritten to give the load resistance that should be used with a given grid bias, with the result that

$$\left.\begin{matrix}\text{Load resistance to be}\\ \text{used with grid bias } E_c\end{matrix}\right\} = R_L = R_p\left[\frac{E_c}{\left(\dfrac{E_B - E_0}{\mu}\right) - E_c} - 1\right] \tag{145}$$

In this equation the signs are chosen so that E_c in the equation is positive when the bias is negative. The effect of varying the load resistance when the grid bias is fixed is shown by dynamic characteristics a, b, and c of Fig. (153). Characteristic a is for the value of resistance called for by Eq. (145) and has a slope such that the operating point O is exactly midway between the point where the dynamic characteristic enters the region of high distortion that exists with low plate currents and the point where the grid becomes positive. Curve b shows the dynamic characteristic when the load resistance is higher than the value called for by Eq. (145), and it is apparent that in this case the grid will go positive with a signal that does not carry the operating point to anywhere near the minimum allowable plate current. Similarly curve c is a dynamic characteristic for a lower load resistance than is called for by Eq. (145), and it is seen that a sine-wave signal having an amplitude sufficient to carry the grid to zero potential during one half cycle will cause the operating point to extend into the high-distortion region during the other half cycle. The full possibilities of the power tube for the particular grid bias shown are therefore realized only in case a. A grid bias that is either higher or lower than that called for by Eq. (144) has much the same effect as varying the load resistance, since the proper load resistance varies with the bias.

---

that the alternating current in the load is $(I_0 - I_{min}) = (I_{max} - I_0)$, while the alternating voltage across the load is $(E_B - E_{min}) = (E'_{max} - E_B)$, where E_{min} and I_{min} are the minimum instantaneous values reached by the plate potential and plate current and E'_{max} and I_{max} are the maximum instantaneous values of plate potential and plate current, respectively. Hence

$$\text{Power output} = \frac{(E_B - E_{min})(I_0 - I_{min})}{2} \tag{142b}$$

$$= \frac{(E'_{max} - E_{min})(I_{max} - I_{min})}{8} \tag{142c}$$

$$\text{Efficiency} = \frac{1}{2}\left(1 - \frac{E_{min}}{E_B}\right)\left(1 - \frac{I_{min}}{I_0}\right) \tag{143b}$$

$$= \frac{1}{8}\left(\frac{E'_{max} - E_{min}}{E_B}\right)\left(\frac{I_{max} - I_{min}}{I_0}\right) \tag{143c}$$

These alternative forms are particularly useful in dealing with pentodes and in circumstances where the grid is driven positive.

The undistorted power output that can be obtained from a power amplifier adjusted to the proper grid bias for the load resistance being used is given by the equations

$$\left.\begin{array}{l}\text{Undistorted power output}\\ \text{with proper grid bias}\end{array}\right\} = (E_B - E_0)^2\frac{R_L}{2(R_L + 2R_p)^2} \qquad (146a)$$

$$= \tfrac{1}{2}(E_B - E_{\min})(I_0 - I_{\min}) \qquad (146b)$$

$$= \tfrac{1}{8}(E'_{\max} - E_{\min})(I_{\max} - I_{\min}) \qquad (146c)$$

where

$E_{\min}$ and $I_{\min}$ = minimum instantaneous plate voltage and current, respectively

$E'_{\max}$ and $I_{\max}$ = maximum instantaneous values of these same quantities

E_B and I_0 = plate voltage and plate current, respectively, at the operating point.

It is to be noted that Eq. (146c) gives the *exact* power output, even taking into account the curvature of the dynamic characteristic, provided only even harmonics are produced, as seen from Eq. (151e).

The load resistance which gives the maximum possible undistorted power output that can be obtained from the tube at a given plate voltage is obtained by differentiating Eq. (146a) with respect to load resistance, equating the result to zero, and solving for load resistance. This operation shows that *the maximum undistorted power output is obtained with a load resistance that equals twice the plate resistance of the tube.* Substituting these relations in Eq. (144) shows that the grid bias for this load resistance should be

$$(\text{Grid bias when } R_L = 2R_p) = \frac{3}{4}\left(\frac{E_B - E_0}{\mu}\right) \qquad (147)$$

Substitution in Eq. (146a) shows that, when the load resistance is twice the plate resistance, the undistorted power output is

$$\text{Maximum possible undistorted power output} = \frac{(E_B - E_0)^2}{16R_p} \qquad (148)$$

The power that the plate-supply voltage must furnish to the tube is equal to the product of plate voltage and plate current at the operating point, and is unchanged by the presence of a signal (assuming that the characteristic curves of the tube are straight lines). When no signal is applied to the grid of the tube, all this energy is dissipated at the plate and appears as heat that must be radiated, while in the presence of a signal the plate loss is reduced by an amount equal to the amplifier output. The ratio of the maximum power output that can be obtained from a tube to the power that must be supplied to the plate from the

voltage-supply source is called the plate efficiency and, when the grid bias has the value called for by Eq. (144), is

$$\text{Plate efficiency} = \frac{\text{maximum power output}}{\text{power supplied to plate}} = \frac{R_L}{2(R_L + 2R_p)} \times \left(1 - \frac{I_{\min}}{I_0}\right)\left(1 - \frac{E_0}{E_B}\right) \quad (149a)$$

$$= \frac{1}{2}\left(1 - \frac{E_{\min}}{E_B}\right)\left(1 - \frac{I_{\min}}{I_0}\right) \quad (149b)$$

$$= \frac{(E'_{\max} - E_{\min})(I_{\max} - I_{\min})}{8E_B I_0} \quad (149c)$$

The notation is the same as previously used. Equation (149*c*) is particularly important as it gives the exact plate efficiency even when the dynamic characteristic is curved, provided the output contains no odd harmonics and provided the plate current I_0 at the operating point is taken in Eq. (149*c*) to be the actual d-c plate current present when the signal is applied, *i.e.*, the sum of the "rectified" plate current given by Eq. (151*a*) or Eq. (152*a*) and the plate current when no signal is applied.

The plate efficiency of a Class A amplifier has a maximum possible value of 50 per cent, but, with tubes operating at plate potentials of 200 to 300 volts and a load resistance approximately twice the plate resistance, the practical efficiency is seen by Eq. (149*a*) to be of the order of 15 to 20 per cent when reasonable values of $I_{\min}/I_0$ and E_0/E_B are selected. This low efficiency comes about as a result of the fact that the maximum plate current $I_{\max}$ flows when the plate potential is at its minimum $E_{\min}$ (as seen from Fig. 152). Hence, when the plate current is high in proportion to the plate-supply voltage, $E_{\min}/E_B$ will necessarily be large. Higher efficiencies, ranging above 30 per cent, can be realized in tubes requiring a high supply voltage in proportion to the rated plate current, since then the desired maximum plate current can be drawn with a value of $E_{\min}$ that is a smaller fraction of the supply voltage E_B (or, what is the same thing, load resistances greater than twice the plate resistance can be used).

Methods of Adjusting Power Amplifiers.—Class A amplifiers may be adjusted in one of several different ways, depending upon the objective desired. It has already been shown that, in order to obtain the maximum possible power output without driving the grid positive during the positive half cycles of the signal and without making the instantaneous plate current less than a fixed value during the negative half cycles, the load resistance should be equal to twice the plate resistance. This is the adjustment usually employed in small power tubes such as those used in radio receivers where the plate-supply voltage E_B is low (250 volts or less).

When the plate-supply voltage exceeds 250 to 300 volts, as is the case with all large power tubes, it is seldom permissible to operate the tube under conditions corresponding to a load resistance equal to twice the plate resistance because this calls for a grid bias that will ordinarily make the plate current exceed the rated value. Under these conditions the allowable plate dissipation is the limiting factor and determines the grid bias that must be employed. This bias is more negative than that called for to obtain the maximum possible power output, and therefore it requires a load impedance greater than twice the plate resistance, as is seen from Eq. (145).

The maximum output in proportion to the signal voltage applied to the grid of the tube is obtained when the load resistance equals the plate resistance. Amplifiers are seldom adjusted for this condition, however, because higher values of load resistance give more undistorted output from the tube, and at the same time require less direct-current plate power. These factors are more important than the slight saving in input signal obtained by making the load and plate resistances equal.

Determination of Exact Dynamic Characteristic and the Calculation of Distortion.[1]—The discussion given above is approximate in that it assumes the dynamic characteristic is a straight line over the operating range. A more exact analysis of amplifier performance including the amount of distortion to be expected can be obtained by using the exact dynamic characteristic to take into account the curvature of the tube characteristics. The procedure for deriving the actual dynamic characteristic is as follows: One point on the dynamic characteristic is the operating point (*i.e.*, the point corresponding to the plate voltage and grid bias that exist with no signal), where the plate voltage and plate current can be represented by E_p' and I_p', respectively. Other points can be determined for a load resistance R_L by noting that the plate current I_p'' at which the dynamic characteristic crosses the E_p-I_p static curve for a plate voltage of E_p'' is given by the relation

$$R_L(I_p'' - I_p') = (E_p' - E_p'') \tag{150}$$

This is equivalent to stating that the voltage drop produced in the load resistance by the difference between the plate current at the operating point and the plate current at the point in question is the amount by which the plate voltage at the point in question differs from the plate voltage at the operating point. By use of Eq. (150) the dynamic charac-

[1] Discussions of the dynamic characteristic and its use in determining distortion are given by E. W. Kellogg, Design of Non-distorting Power Amplifiers, *Trans. A.I.E.E.*, vol. 44, p. 302, 1925; J. C. Warner and A. V. Loughren, The Output Characteristics of Amplifier Tubes, *Proc. I.R.E.*, vol. 14, p. 735, December, 1926.

teristic can be calculated point by point for a resistance load, with a result such as is shown in Fig. 152.

The dynamic characteristic gives the relationship between the actual plate current and the voltage applied to the grid, and so can be used to determine the wave shape of the amplifier output, as shown in Fig. 152. In triode Class A amplifiers operated so that the grid does not go positive and so that the minimum instantaneous plate current does not reach zero, the distortion that is produced when a sine-wave voltage is applied to the grid consists primarily of a second harmonic component of plate current plus a direct-current (or "rectified") component that causes the average plate current to increase when the signal is applied. These components, as well as the fundamental frequency output can be determined from a knowledge of the instantaneous plate current when the applied voltage is at zero, maximum and minimum according to the formulas

$$\text{Direct-current component} = A_0 = \frac{I_{\max} + I_{\min} - 2I_0}{4} \tag{151a}$$

$$\text{Fundamental} = A_1 = \frac{I_{\max} - I_{\min}}{2} \tag{151b}$$

$$\text{Second harmonic} = A_2 = \frac{I_{\max} + I_{\min} - 2I_0}{4} \tag{151c}$$

$$\frac{\text{Second harmonic}}{\text{Fundamental}} = \frac{A_2}{A_1} = \frac{I_{\max} + I_{\min} - 2I_0}{2(I_{\max} - I_{\min})} \tag{151d}$$

$$\text{Power output} = \frac{A_1{}^2 R_L}{2} = \frac{(I_{\max} - I_{\min})^2 R_L}{8} = \frac{(E'_{\max} - E_{\min})(I_{\max} - I_{\min})}{8} \tag{151e}$$

$$\text{Plate efficiency} = \frac{A_1{}^2 R_L}{2E_B(I_0 + A_0)} \tag{151f}$$

The notation is as indicated in Fig. 152, and as in Eq. (146).

When third- and fourth-harmonic components can be expected in the output, as will be the case with pentodes and with badly overloaded triodes, one can determine the components by obtaining from the dynamic characteristic the instantaneous plate current when the applied voltage wave is at zero, maximum, and minimum, and at 0.707 of the maximum and minimum values, as shown in Fig. 152. Then[1]

$$\text{Direct-current component} = A_0 = \frac{\tfrac{1}{2}(I_{\max} + I_{\min}) + I_2 + I_3 - 3I_0}{4} \tag{152a}$$

[1] See G. S. C. Lucas, Distortion in Valve Characteristics, *Exp. Wireless and Wireless Eng.*, vol. 8, p. 595, November, 1931.

$$\text{Fundamental} = A_1 = \frac{\sqrt{2}(I_2 - I_3) + I_{\max} - I_{\min}}{4} \tag{152b}$$

$$\text{Second harmonic} = A_2 = \frac{I_{\max} + I_{\min} - 2I_0}{4} \tag{152c}$$

$$\text{Third harmonic} = A_3 = \frac{I_{\max} - I_{\min} - 2A_1}{2} \tag{152d}$$

$$\text{Fourth harmonic} = A_4 = \frac{2A_0 - I_2 - I_3 + 2I_0}{2} \tag{152e}$$

$$\text{Power output} = \frac{A_1{}^2R_L}{2} \tag{152f}$$

$$\text{Plate efficiency} = \frac{A_1{}^2R_L}{2E_B(I_0 + A_0)} \tag{152g}$$

By the use of these formulas it is possible to get the exact performance, including the power output and the distortion for any operating point and any load resistance.

In using Eqs. (151) and (152) one needs only the plate current at certain critical parts of the cycle, and does not require the entire dynamic characteristic. These critical points can be obtained most readily by the construction shown in Fig. 154 rather than by plotting out the entire characteristic as in Fig. 152. Here a line called the load line is drawn through the operating point and the point $E_p = O$, $I_p = \frac{E_B}{R_L} + i_0$, as shown in Fig. 154. Consideration shows that, as the grid potential is varied by the signal, the instantaneous plate current and plate voltage must vary along this load line, with the point for any particular signal voltage on the grid being where the load line intersects the corresponding grid potential line.

The dynamic characteristic is very helpful in analyzing distortion, but it can be used only when the signal is a simple sine wave and the load is a resistance. Furthermore, in the derivation of the dynamic characteristic it was assumed that the load offered the same resistance to all components contained in the output current. Under practical conditions the load is very nearly the same for all alternating components, but the d-c current is usually by-passed around the load resistance through a transformer primary and so produces no voltage drop in the load. This causes the direct-current voltage actually appearing at the plate of the tube to be the same when the signal voltage is applied as when there is no signal, whereas, according to the dynamic characteristic the direct-current voltage should be reduced slightly. This effect is commonly neglected, although it must be taken into account if maximum accuracy is to be obtained.

When the load impedance has a reactive component, the dynamic characteristic becomes an ellipse as shown in Fig. 155. The analysis by the dynamic characteristic then becomes so involved, even when the tube characteristics are assumed to be straight lines, as to have little practical value.[1]

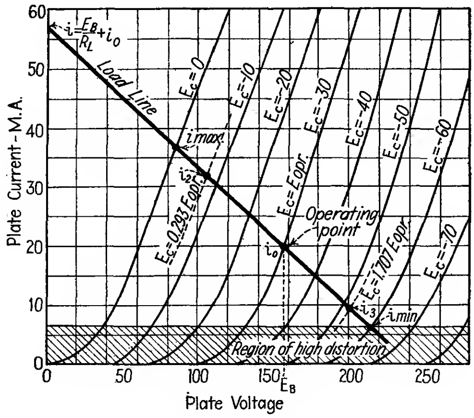


FIG. 154.—Load line drawn on E_p-I_p characteristic curves, showing critical points used in Eqs. (151) and (152) to determine distortion.

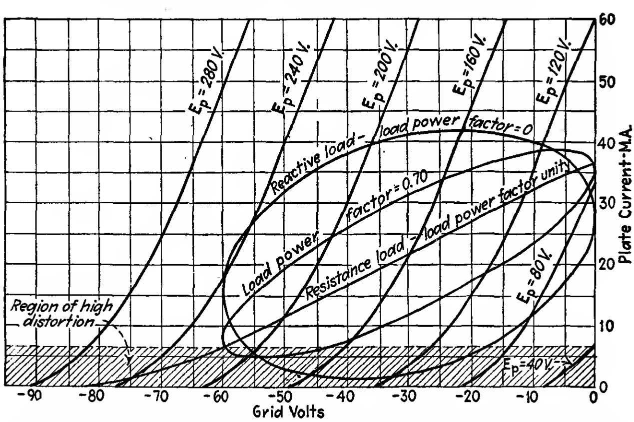


FIG. 155.—Dynamic characteristics for load impedances of varying power factors but constant magnitude, showing how the reactive loads cause the characteristic to open up into ellipses that extend to a lower value of plate current than are reached with the same signal when the load is a resistance.

The limitations of the dynamic characteristics can be avoided by using the power-series method of analyzing the behavior of the power amplifier, as discussed in Sec. 55. In this way it is possible to determine the exact behavior with any type of load and with complex signal volt-

[1] For an analysis of power amplifiers with reactive loads, see Manfred von Ardenne, On the Theory of Power Amplification, *Proc. I.R.E.*, vol. 16, p. 193, February, 1928.

ages. The disadvantage of the power-series method is that it requires a knowledge of the second and higher order derivatives of the tube characteristic, and these can be obtained only from laboratory tests. In contrast with this, the dynamic characteristic method of analysis requires merely a set of characteristic curves such as published by every tube maker.

It will be noted that, when a signal is applied to a power amplifier, a d-c "rectified" current is produced which adds to the normal plate current. This added direct current increases the direct-current power supplied to the tube when a signal is applied and therefore must be taken into account in accurate calculations of efficiency. It also shifts the grid bias in self-bias arrangements, which must therefore be proportioned with this effect in mind.

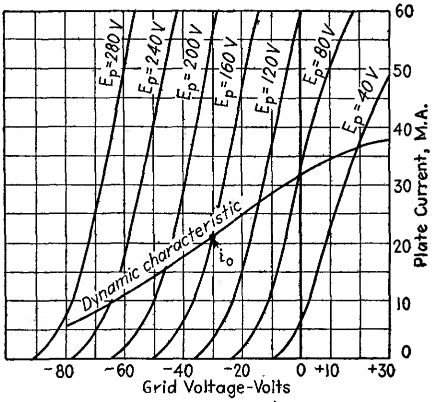


FIG. 156.—Characteristic curves of power-amplifier tube, with dynamic characteristic extending into positive grid region.

It can be shown from the power-series method of analysis that, when the distortion is primarily of the second order (as is the case with triodes), the amount of second-order distortion is related to the "rectified" d-c plate current. Thus, when the signal is a sine wave, Eq. (135a) shows that the equivalent voltage producing the "rectified" direct current is the same as the crest alternating-current second-harmonic voltage, which can be considered as acting in the plate circuit to produce the second-order effects. If the "rectified" current is ΔI and the load impedance to direct current is negligible, then the equivalent direct current and crest second-harmonic voltages that can be considered as acting in the plate circuit are $R_p\Delta I$. The equivalent signal voltage is μE_s, so that the percentage of second-harmonic distortion is $\frac{R_p \Delta I}{\mu E_s}100 = \frac{\Delta I}{G_m E_s}$, and can be readily obtained by observing changes in plate current with a direct-current meter.

Class A Power Amplifiers Operated So That the Grid Is Driven Positive. The dynamic characteristic of ordinary tubes is found to be substantially linear up to grid voltages that are appreciably positive, as shown in Fig. 156, and by taking advantage of this it is possible to increase the plate efficiency and power output obtainable from a given tube. The improvement in plate efficiency comes about as a result of the fact that, when the grid is driven positive, the desired maximum plate current can be drawn

with a much lower plate potential E_{min} than if the grid is not positive. Since the operating point and hence the plate power input can be the same when the grid is driven positive as when it is not, the higher efficiency increases the power output correspondingly. With small tubes which must be operated with a load resistance approximately twice the plate resistance with negative grid operation, positive grid operation will make it possible to increase the plate efficiency to about 35 per cent, and so roughly doubles the undistorted power output.

When the grid of the power tube is driven positive, it is necessary to adjust the grid bias and load impedance so as to maintain a proper balance with respect to the plate-supply voltage, plate resistance, amount the grid is driven positive, and the minimum allowable instantaneous plate current. When the restriction that E_s equals E_c in Eqs. (144) to (146) and (149) is removed, one has:[1]

$$\left.\begin{matrix}\text{Proper negative grid bias for}\\ \text{load resistance } R_L\end{matrix}\right\} = \left(\frac{E_B - E_0}{\mu}\right)\left(\frac{R_L + R_p}{R_L + 2R_p}\right) - E_{G_{max}}\frac{R_p}{R_L + 2R_p} \tag{153a}$$

$$\left.\begin{matrix}\text{Load resistance to be used}\\ \text{with grid bias } E_c\end{matrix}\right\} = R_p\left[\frac{E_c + E_{G_{max}}}{\left(\dfrac{E_B - E_0}{\mu}\right) - E_c} - 1\right] \tag{153b}$$

$$\left.\begin{matrix}\text{Undistorted power output}\\ \text{with proper grid bias}\end{matrix}\right\} = (E_B - E_0 + \mu E_{G_{max}})^2\frac{R_L}{2(R_L + 2R_p)^2} \tag{154a}$$

$$= \tfrac{1}{2}(E_B - E_{min})(I_0 - I_{min}) \tag{154b}$$

$$= \frac{(E'_{max} - E_{min})(I_{max} - I_{min})}{8} \tag{154c}$$

$$\text{Plate efficiency} = \frac{\text{maximum power output}}{\text{power supplied to plate}} = \frac{R_L}{2(R_L + 2R_p)} \times \left(1 - \frac{E_0}{E_B} + \frac{\mu E_{G_{max}}}{E_B}\right)\left(1 - \frac{I_{min}}{I_0}\right) \tag{155a}$$

$$= \frac{1}{2}\left(1 - \frac{E_{min}}{E_B}\right)\left(1 - \frac{I_{min}}{I_B}\right) \tag{155b}$$

$$= \frac{1}{8}\left(\frac{E'_{max} - E_{min}}{E_B}\right)\left(\frac{I_{max} - I_{min}}{I_0}\right) \tag{155c}$$

The notation is the same as in Eqs. (141) to (149), with the addition that $E_{G_{max}}$ is the amount the grid is driven positive ($E_{G_{max}}$ equals crest value of signal voltage minus grid bias) and is not to be confused with the maximum instantaneous plate potential E'_{max}. It is to be noted that Eq. (154c) gives the correct output when even harmonic distortion is present

[1] The derivation of these equations is readily obtained from the analysis given in the footnote on page 278.

[compare with Eq. (151*e*)], and that Eq. (155*c*) likewise is correct in the presence of even harmonic distortion provided the current I_0 in the equation is interpreted to be the sum of the rectified current A_0 given by Eq. (151*a*) plus the current actually at the operating point when no signal is applied.

The disadvantage of driving the grid positive is that this causes grid current to flow at the positive peaks of the exciting voltage, flattening these peaks and thereby introducing distortion. The factors involved in this distortion can be understood by using Thévenin's theorem to represent the exciting source as a generator of potential E_s having an internal impedance Z_s, where E_s is the voltage that would appear at the grid terminals if the grid drew no current and Z_s is the impedance that the grid of the tube sees when looking toward the exciting voltage. If the instantaneous grid current at the peak when the grid is most positive is ΔI_g, then the voltage drop in the generator impedance is $Z_s \Delta I_g$ and the actual exciting voltage at the peak of the cycle is $(E_s - Z_s \Delta I_g)$ instead of the undistorted value of E_s. It is therefore apparent that, in order to keep the distortion small, the voltage drop $Z_s \Delta I_g$ must be made small compared with the crest exciting voltage E_s by making the source impedance Z_s low or by not driving the grid far enough positive to draw much grid current, or both. Assuming that grid current flows for less than one-fourth of the cycle, it can be found from Eq. (152) that the distortion of the driving voltage produced by the flattening of the positive peaks is:

$$\left.\begin{aligned}\frac{\text{Second harmonic}}{\text{Fundamental}} &= \frac{Z_s \Delta I_g}{4E_s - Z_s \Delta I_g} \sim \frac{Z_s \Delta I_g}{4E_s}\\ \frac{\text{Third harmonic}}{\text{Fundamental}} &= \frac{Z_s \Delta I_g}{4E_s - Z_s \Delta I_g} \sim \frac{Z_s \Delta I_g}{4E_s}\\ \frac{\text{Fourth harmonic}}{\text{Fundamental}} &= \frac{Z_s \Delta I_g}{8E_s - 2Z_s \Delta I_g} \sim \frac{Z_s \Delta I_g}{8E_s}\end{aligned}\right\} \tag{156}$$

The phase of the second harmonic is such as to tend to cancel the second harmonic generated by the curvature of the dynamic characteristic. An idea of the amount of flattening permissible can be gained by noting that, if the voltage drop produced by the grid-current in flowing through the source impedance is 20 per cent of the crest signal voltage, second- and third-harmonic components are each 5 per cent and the fourth is 2.5 per cent.

Tubes in which the grid is driven positive are preferably operated from transformer-coupled amplifiers having a low voltage step-up, or even a step-down, so that the impedance of the source as viewed by the grid of the power tube will be low. The low transformation ratio reduces the voltage gain of the driving amplifier, and this together with the distortion of the

TABLE VIII.—CHARACTERISTICS OF REPRESENTATIVE SMALL POWER TUBES

Type	Description	Cathode data			Use*	Typical operating conditions									
		E_f, volts	I_f, amp.	Type		E_p, volts	E_{sg}, volts	E_{cg},† volts	I_p, ma	I_{sg}, ma	R_p, ohms	μ	G_m, μ mho	Load resistance, ohms	Power output, watts
31	Triode	2	0.13	Filament	Class A	180	...	−30	12.3	...	3,600	3.8	1,050	5,700	0.375
45	Triode	2.5	1.5	Filament	Class A	250	...	−50	34	...	1,610	3.5	2,175	3,900	1.60
					Class AB, fixed bias	275	...	−68	28	...	....	.....	.....	3,200	18
					Class AB, self-bias	275	...	775 Ω	72	...	....	.....	.....	5,060	12
2A3	Triode	2.5	2.5	Filament	Class A	250	...	−45	60	...	800	4.2	5,250	2,500	3.5
					Class AB, fixed bias	300	...	−62	80	...	....	.....	.....	3,000	15
					Class AB, self-bias	300	...	780 Ω	80	...	....	.....	.....	5,000	10
19	Twin triode	2	0.26	Filament	Class B	135	...	0	10	...	....	.....	.....	10,000	2.1
53	Twin triode	2.5	2	Heater	Class B	300	...	0	35	...	....	35	.....	10,000	10
46	Dual grid	2.5	1.75	Filament	Class B	400	...	0	12	...	....	.....	.....	5,800	20
					Class A, triode	250	...	−33	22	...	2,380	5.6	2,350	6,400	1.25
49	Dual grid	2	0.120	Filament	Class A, triode	135	...	−20	6	...	4,175	4.7	1,125	11,000	0.170
					Class B	180	...	0	4	...	....	.....	.....	12,000	3.5
59	Triple grid	2.5	2	Heater	Class A, triode	250	...	−28	26	...	2,300	6	2,600	5,000	1.25
					Class A, pentode	250	250	−18	35	9	40,000	100	2,500	6,000	3
					Class B, triode	400	...	0	26	...	....	.....	.....	6,000	20
89	Triple grid	6.3	0.4	Heater	Class A, triode	250	...	−31	32	...	2,600	4.7	1,800	5,500	0.90
					Class A, pentode	250	250	−25	32	5.5	70,000	125	1,800	6,750	3.4
					Class B, triode	180	...	0	6	...	....	.....	.....	9,400	3.5

Type	Description	Cathode data			Use*	Typical operating conditions									
		E_f, volts	I_f, amp.	Type		E_p, volts	E_{sg}, volts	E_{cg},† volts	I_p, ma	I_{sg}, ma	R_p, ohms	μ	G_m, μ mho	Load resistance, ohms	Power output, watts
33	Pentode	2.0	0.26	Filament	Class A	180	180	−18	22	5	55,000	90	1,700	6,000	1.4
2A5	Pentode	2.5	1.75	Heater	Class A, triode	250	. . .	−20	31	. . .	2,700	6.2	2,300	3,000	0.65
					Class A, pentode	250	250	−16.5	3.	6.5	80,000	190	2,350	7,000	3
					Class AB, triode fixed bias	350	. . .	−38	45	. . .	....	.....	. . . .	6,000	18
					Class AB, triode self-bias	350	. . .	730 Ω	50	. . .	....	.....	.....	10,000	14
6F6	Pentode	6.3	0.7	Heater	Class A, triode	250	. . .	−20	31	. . .	2,600	7	2,700	4,000	0.85
					Class A, pentode	315	315	−22	42	8	75,000	200	2,650	7,000	5
					Class AB, pentode fixed bias	375	250	−26	34	5	....	.....	.....	10,000	19
					Class AB, pentode self-bias	375	250	340 Ω	54	8	....	.....	.....	10,000	19
6L6	Beam tube	6.3	0.9	Heater	Class A	375	250	−17.5	57	2.5	22,500	135	6,000	4,000	11.5
					Class AB₁ (no I_g)	400	300	−25	102	6	....	.....	.....	6,600	34
					Class AB₂ (with I_g)	400	300	−25	102	6	....	.....	.....	3,800	60

Note: Triode connection for pentode tubes with internally connected suppressor consists in connecting screen and plate together. The resulting characteristics are essentially those of a triode in spite of the suppressor grid.

* Where "Use" is Class AB or Class B, the output given is for two tubes and the load resistance given is the proper value from plate to plate, while the plate current is the zero-signal value for the two tubes.

† Where self-bias is used, the bias resistance is given.

exciting voltage is the price paid for the increased output. Resistance coupling to a grid which is to be driven positive is not desirable (unless the grid leak is replaced by a choke as in Fig. 114) because the grid current in flowing through the grid leak produces a bias on the grid of the power tube.

The proper use of negative feedback (see Sec. 52) is of considerable assistance in reducing distortion of the driving voltage as a result of grid current.

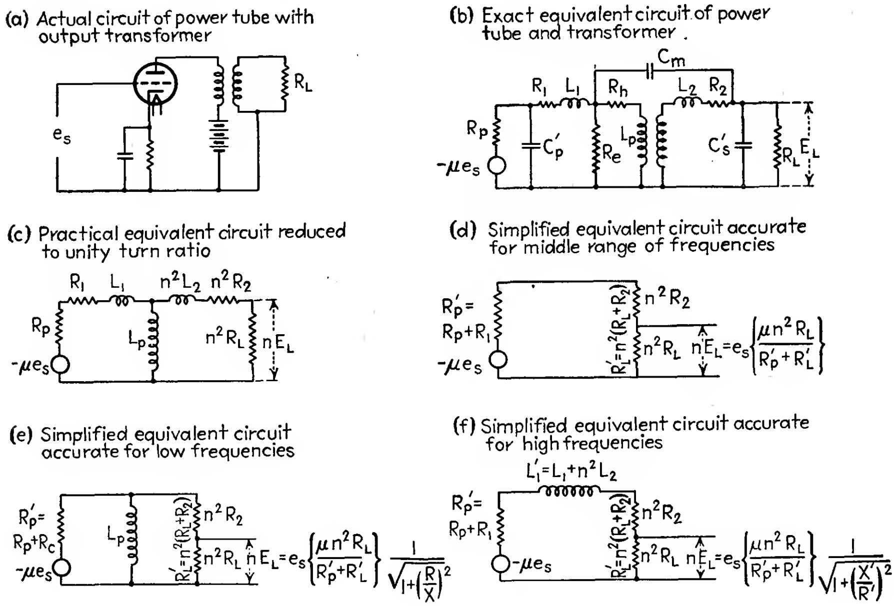
Tubes for Power Amplification.—The type of tube best suited for power amplification differs in its characteristics from the tube best adapted to voltage amplification. Since the plate efficiency of an ideal tube, as given in Eqs. (149) and (155), depends only upon the ratio of load to plate resistance and not upon the plate voltage or current, it is apparent that the undistorted output available from a power tube is approximately proportional to the plate power. This means that a tube having a high power capacity must be operated at a high plate voltage or at a high plate current, or both.

In the smaller power tubes, such as those employed in radio receivers, the plate voltage available is limited by practical requirements to 250 to 300 volts, and a large power output can be obtained only by increasing the plate current at this fixed plate voltage. This calls for a cathode with high electron emission and an amplification factor that is low enough to enable the limited plate voltage to produce the relatively strong electrostatic field in the vicinity of the cathode which is needed to draw the large space current. The usual operating condition under such circumstances is with a load resistance approximately twice the plate resistance. A tube with a low amplification factor requires a large grid bias, as indicated by Eq. (144), and this means that a large signal voltage must be applied to the grid to develop the full output of the tube. This disadvantage represents the price paid to obtain a large power output with a limited plate voltage.

In large power tubes, particularly those operating at plate potentials of 500 volts and higher, the tubes are designed to operate with a much higher ratio of plate voltage to plate current than are low-voltage tubes, and the limiting factor in operation becomes plate dissipation. With such tubes, the grid bias is fixed by the allowable plate current, and the load impedance is selected accordingly, with load resistances greater than twice the plate resistance normally required. The result is higher plate efficiency than with lower voltage power tubes and also less exciting voltage because the high plate voltage and low plate current permit the use of a higher amplification factor.

The characteristics of typical Class A power-amplifier tubes of the sizes used in radio receivers and small public-address systems are tabulated in Table VIII. All these tubes are characterized by a high electron

emission, low amplification factor, and, except in the case of the 2A3, an optimum operating condition corresponding to a load resistance approximately twice plate resistance. The tubes included in Table VIII are



μ = amplification factor of tube
R_p = plate resistance of tube (in a push-pull amplifier R_p is twice the plate resistance of one tube)
R_1 = d-c resistance of primary winding
R_p' = $R_p + R_1$ = effective plate resistance
R_2 = d-c resistance of secondary winding
R_L = load resistance
R_L' = effective load resistance reduced to unity turn ratio
R = $R_L'R_p'/(R_L' + R_p')$ – resistance formed by R_p' and R_L' in parallel
R' = $R_L' + R_p'$ = sum of effective load and effective plate resistances
R_h = resistance representing hysteresis loss
R_e = resistance representing eddy-current loss
n = step down voltage ratio = ratio of primary to secondary turns
L_p = primary inductance with appropriate d-c saturation
L_1 = leakage inductance of primary winding
L_2 = leakage inductance of secondary winding
L_1' = $L_1 + n^2L_2$ = total leakage inductance reduced to unity turn ratio
C_m = capacity between primary and secondary windings of transformer
C_s' = Secondary distributed capacity
C_p' = Capacity in shunt with primary (tube plus transformer capacity)
X = ωL_p = reactance of transformer primary inductance
X' = $\omega L_1'$ = reactance of transformer leakage inductance
e_s = input voltage
E_L = output voltage

FIG. 157.—Actual circuit of power amplifier with output transformer, together with equivalent circuits used in determining the frequency response.

triodes, double-grid tubes, triple-grid tubes, and pentodes. The double- and triple-grid tubes can be connected as Class A triode amplifiers by using the inner grid as the control electrode and connecting the remaining grid or grids to the plate. Pentode tubes are considered in Sec. 58. Data on large triode tubes used as Class A amplifiers in large public-address

TABLE IX.—CHARACTERISTICS OF REPRESENTATIVE TRANSMITTING POWER TUBES

Triode Tubes

Type	Filament			Method of cooling	μ	Allowable plate loss, watts	Typical operating conditions as audio amplifier										Typical operating conditions as Class C amplifier							
	E_f, volts	I_f, amp.	Type				Use*	Plate voltage	Control-grid voltage	Peak exciting voltage per tube	Plate* current, ma	Mutual conductance, μ mhos	Plate resistance, ohms	Load resistance, ohms	Undistorted output, watts*	Plate efficiency, per cent	Plate voltage	Control-grid voltage	Peak signal voltage	Plate current, ma	Grid current, ma	Driving power, watts	Output, watts	Plate efficiency, per cent
204A	11	3.85	Thoriated	Air	23	250	Class B	3,000	−100	250	80	.....	.....	20,000	700	62.7	2,500	−200	440	250	30	15	450	72
207	22	52	Tungsten	Water	20	7,500	Class B	12,500	−575	1,150	400	.....	.....	10,000	22,500	64.4	12,000	−1,600	2,650	1,670	90	235	15,000	75
211	10	3.25	Thoriated	Air	12	100	Class A Class B	1,250 1,250	−80 −100	75 205	60 20	3,300	3,600	9,200 9,000	19.7 260	26.3 65	1,250	−225	375	150	18	7	130	69.3
800	7.5	3.25	Thoriated	Air	15	35	Class B	1,250	−70	150	30	.....	.....	21,000	106	65	1,000	−135	260	70	15	3	50	71.5
801	7.5	1.25	Thoriated	Air	8	20	Class A Class B	600 600	−55 −75	50 160	30 8	1,840	4,300	7,800 10,000	3.8 45	21.2 57.7	600	−150	260	65	15	4	25	64
848	22	52	Tungsten	Water	8	7,500	Class A Class B	8,000 12,500	−730 −1,560	800 2,080	900 400	....	....	5,200 10,000	2,000 22,000	27.8 62.8	10,000	−2,000	2,900	1,450	100	310	10,000	69
851	11	15.5	Thoriated	Air	20.5	600	Class A Class B	2,500 3,000	−92 −135	87 245	240 110	....	1,600	5,000 5,600	160 2,400	26.7 66.6	2,500	−250	450	900	100	45	1,700	75.5
852	10	3.25	Thoriated	Air	12	100	Class B	3,000	−250	390	14	.....	.....	41,000	320	66.5	3,000	−600	850	85	15	12	165	64.8
862	33	207	Tungsten	Water and forced air	48	100,000	Class B	12,000	0	1,000	3,000	.....	.....	1,800	90,000	57.5	18,000	−1,000	2,550	8,333	900	2,400	100,000	66.7

* Where "Use" is Class B, power output is for the two tubes in push-pull, load resistance is resistance from plate to plate, and plate current is zero-signal current for two tubes.

Screen-grid and Pentode Tubes

Type	Description	Filament			Allowable plate loss, watts	Allowable screen loss, watts	Connection	Typical operating conditions as Class C amplifier										
		E_f, volts	I_f, amp.	Type				Plate voltage	Control-grid voltage	Screen-grid voltage	Suppressor voltage	Plate, current ma	Screen-grid current, ma	Control-grid current, ma	Peak signal voltage	Driving power, watts	Output, watts	Plate efficiency
802	Pentode	6.3	0.9	Heater	10	6	Pentode Tetrode	500 500	−100 −60	200 100	0 ..	45 45	22 15	6 6	155 90	0.9 0.5	14 12	62.1 53.3
803	Pentode	10	5	Thoriated	125	30	Pentode	2,000	−90	500	40	160	45	12	175	2	210	65
804	Pentode	7.5	3	Thoriated	40	15	Pentode Tetrode	1,250 1,250	−100 −100	300 180	45 ..	92 92	27 23	7 8	150 160	0.9 1.2	80 80	69.5 69.5
860	Screen grid	10	3.25	Thoriated	100	10	....	3,000	−150	300	..	85	..	15	...	7	165	64.7
861	Screen grid	11	10	Thoriated	400	35	....	3,500	−250	500	..	300	..	40	725	30	700	66.7
865	Screen grid	7.5	2	Thoriated	15	3	....	750	−80	125	..	40	..	5.5	...	1	16	53.4

systems and in radio transmitters are given in Table IX. These tubes are characterized by a high plate voltage in proportion to plate current and a relatively high amplification factor. The proper operating point is determined by the allowable dissipation, and normal operation is with a load resistance considerably greater than twice the plate resistance.

57. Output Transformers in Class A Amplifiers.—The load impedance is usually coupled to the tube of a Class A power amplifier by means of a transformer as shown in Fig. 157*a*. This arrangement avoids passing the d-c plate current through the load impedance, and also makes it possible by the use of the proper turn ratio to make any load present the desired impedance to the tube.

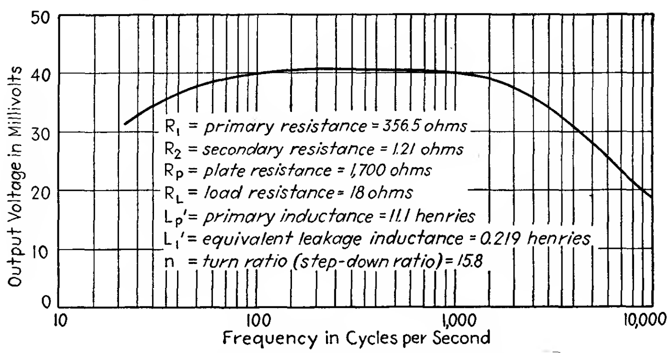


FIG. 158.—Way in which the amplification varies with frequency in a typical power amplifier with output transformer.

The use of an output transformer causes the output of the amplifier to fall off at high and low frequencies in the manner shown in Fig. 158. The falling off at low frequencies is caused by the inductance of the transformer primary, while the falling off at high frequencies is caused by the leakage inductance.

Equivalent Circuits of Amplifier with Output Transformer, and Analysis of the Frequency-response Characteristics.[1]—A quantitative analysis of the frequency-response characteristic of an output transformer can be made by setting up the equivalent circuit of the amplifier tube. The exact equivalent circuit is shown in Fig. 157*b*. This can be reduced under ordinary conditions to the practical equivalent circuit of Fig. 157*c* by taking advantage of the facts that the eddy current and hysteresis losses are usually quite low and that the electrostatic capacities of the transformer can be neglected because of the relatively low plate and load resistances with which they are associated. Further simplification of the equivalent circuit of the transformer is possible by considering restricted ranges of frequencies at one time.

[1] This discussion follows that in F. E. Terman and R. R. Ingebretsen, Output Transformer Response, *Electronics*, vol. 9, p. 30, January, 1936.

In the middle range of frequencies the reactance of the primary inductance is so much greater than the plate and load resistances as to have negligible shunting effect, whereas the frequency is still low enough so that the reactance of the leakage inductance is relatively small. Under these conditions, the equivalent circuit reduces to the network of resistances shown in Fig. 157*d*. Analysis of the voltage and current relations of this circuit gives

$$E_L = e_s\left(\frac{\mu n R_L}{R_p' + R_L'}\right) \tag{157}$$

The notation is shown in Fig. 157.

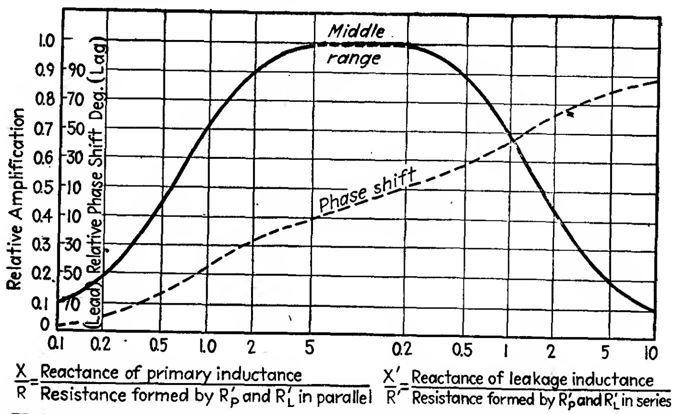


FIG. 159.—Universal amplification curve, showing the way in which the amplification of an output transformer falls off at high and low frequencies.

At low frequencies the shunting effect of the primary inductance must be considered, so that the equivalent circuit takes the form shown in Fig. 157*e*. Analysis of the voltage and current relations of this circuit shows that at low frequencies

$$E_L = e_s\left(\frac{\mu n R_L}{R_p' + R_L'}\right)\frac{1}{\sqrt{1 + (R/X)^2}} \tag{158}$$

It is to be noted that the falling off in response at low frequencies depends only upon the ratio X/R, where X is the reactance of the primary inductance and R is the resistance of the effective plate resistance shunted by the effective load resistance. The way in which the relative amplification depends upon X/R is plotted in Fig. 159, which can be used to determine the low-frequency response. *It is to be noted that the output voltage falls off to 70.7 per cent of its middle-frequency range value when the frequency is such that the reactance of the primary inductance equals the resistance formed by the effective plate resistance R_p' in parallel with the effective load resistance R_L'.*

At high frequencies the shunting effect of the primary inductance can be neglected, but the leakage reactance must be considered, so that the equivalent circuit then takes the form shown at Fig. 157*f*. Analysis of this circuit shows that at high frequencies

$$E_L = e_s\left(\frac{\mu n R_L}{R_p' + R_L'}\right)\frac{1}{\sqrt{1 + (X'/R')^2}} \tag{159}$$

It is to be noted that the falling off in response at high frequencies is determined by X'/R', where X' is the reactance of the leakage inductance and R' is the sum of the effective load and plate resistances. The relation between the relative high-frequency response and X'/R' is plotted in Fig. 159, which can be used to determine the way in which the response falls off at high frequencies. *It will be observed that the output voltage at high frequencies falls to* 70.7 *per cent of its value in the middle range of frequencies when the frequency is such that the reactance of the leakage inductance is equal to the resistance formed by the effective plate resistance R_p' and the effective load resistance R_L' in series.*

A consideration of the above equations and Fig. 159 shows that, in order to obtain good reproduction at low frequencies, the transformer should have a high primary inductance, while, in order to maintain the response to high frequencies, the transformer should at the same time have a low leakage inductance. With any given load and plate resistances, the ratio of highest to lowest frequency satisfactorily reproduced is determined by the ratio of leakage to primary inductance, which can be conveniently termed the *leakage coefficient.* On the other hand, with a given plate resistance and a given transformer, increasing the load resistance will greatly improve the high-frequency response while adversely affecting to a slight extent the response at low frequencies. These effects are illustrated in Fig. 160, which shows the results that can be expected when the primary inductance, the leakage inductance, and the load resistance are varied in turn.

Transformation Ratio and Transformer Efficiency.—It is well-known that placing a load impedance Z_L on the secondary side of a transformer is equivalent to placing another impedance n^2Z_L across the primary terminals, where n is the ratio of primary to secondary turns. When transformer coupling is employed with a power amplifier, the usual practice is to make the turn ratio such that the equivalent primary impedance n^2R_L is equal to the load resistance with which the power tube is designed to operate. This procedure is exactly correct in the case of an ideal transformer having zero primary and secondary copper resistance and it is approximately correct with practical transformers having copper losses.

The output transformer always delivers less energy to the load than is developed by the tube as a result of power losses in the transformer. Most of the transformer loss is caused by the resistance of the windings since flux densities are usually so low that the core losses are negligible.

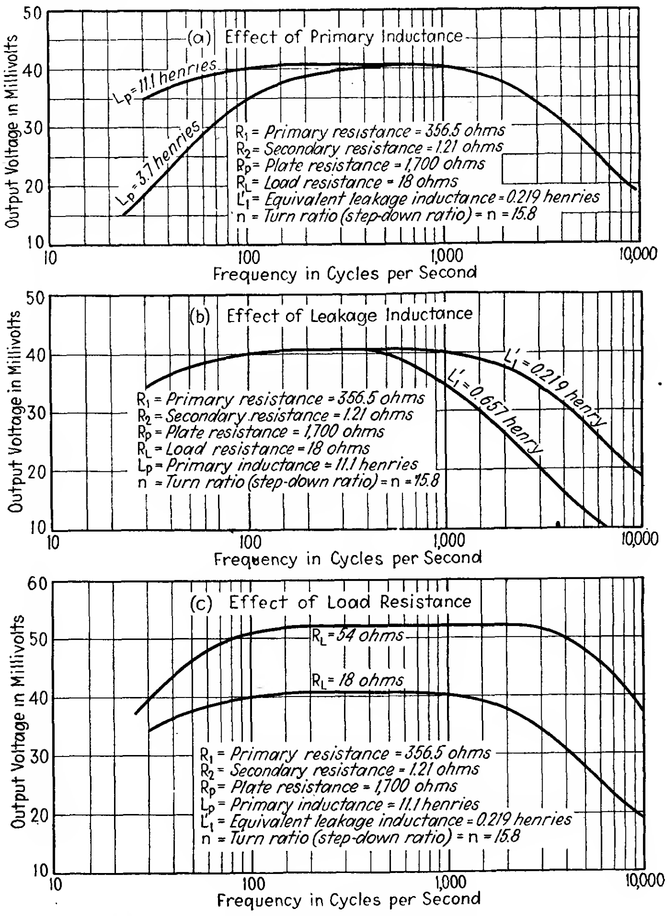


FIG. 160.—Effect on the frequency-response characteristic of varying the circuit constants of an output transformer.

To the extent that the wire resistance accounts for all the transformer dissipation, reference to Fig. 157*d* gives

$$\text{Transformer efficiency} = \frac{\text{power delivered to load}}{\text{power delivered to transformer}} = \frac{1}{1 + \dfrac{R_1 + n^2R_2}{n^2R_L}} \tag{160}$$

Practical efficiencies are of the order of 80 per cent.

Measurement of Transformer Characteristics.—The essential characteristics of an output transformer are the primary inductance, the leakage inductance reduced to unity turn ratio, the turn ratio, and the copper-loss resistances also reduced to unity turn ratio. The primary inductance that is effective is the incremental inductance discussed in Sec. 6; this is the value obtained with the appropriate direct-current saturation in the core and with relatively low alternating flux density. The leakage inductance reduced to unity turn ratio can be determined by short-circuiting the secondary of the transformer and measuring the equivalent inductance appearing across the primary terminals. This inductance depends upon flux paths that are largely in air, and so is unaffected by direct-current saturation, alternating flux density, etc. The resistances of the

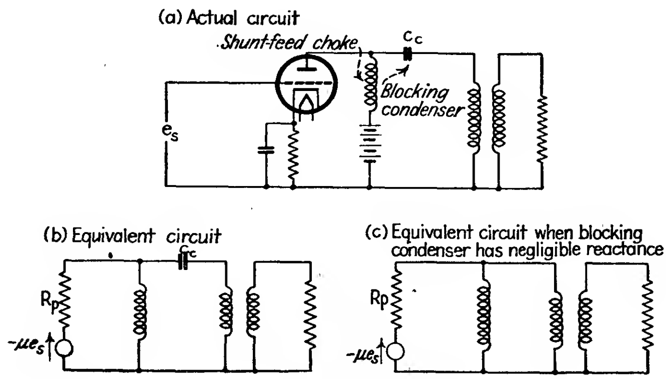


FIG. 161.—Output transformer with shunt feed, together with equivalent circuit.

windings can be measured on a Wheatstone bridge and the turn ratio can be obtained either from the manufacturer or by applying a known voltage to the primary and reading the resulting open-circuit secondary voltage with a vacuum-tube voltmeter.

Transformers with Shunt Feed.—The shunt-feed arrangement illustrated in Fig. 161*a* is often used in order to avoid passing the d-c plate current through the primary winding of the transformer. This eliminates direct-current saturation of the core, thereby increasing the incremental inductance of the primary and making it permissible to reduce the wire size in the transformer primary.

The equivalent circuit of the shunt-fed output transformer is shown in Fig. 161*b*. If the blocking condenser C_c is large enough to be a virtual short circuit, as is usually the case, this reduces to Fig. 161*c*, which is seen to be essentially the same as the equivalent circuit of an ordinary transformer-coupled amplifier if the effective primary inductance is taken as the actual inductance of the transformer primary in parallel with the

shunt-feed choke. By making this change, Eqs. (157), (158), and (159) can be used, and no new analysis is necessary.

The use of shunt feed makes possible considerable improvement in frequency response, and also in the case of very large tubes may reduce the cost of obtaining a given response characteristic. The shunt-feed inductance contributes nothing to the leakage, and so can be made physically large enough to have a high incremental inductance. At the same time, removing the direct current from the primary of the transformer makes it possible to reduce the air gap in the transformer core, which increases the primary inductance without changing the leakage, or makes it possible to obtain a given primary inductance with less leakage inductance.

Shunt feed must be employed with output transformers having high-permeability alloy cores, since the incremental permeability of all such alloys is greatly reduced with even small direct-current magnetization.

Power Rating of Output Transformers.—The maximum current that an output transformer can carry is determined by the heating of the windings, while the maximum voltage that can be applied is limited by the permissible flux density in the core. These two factors operating together determine the power rating of the output transformer.

If the alternating flux density in the core is high, the inductance varies during the cycle because of the non-linear character of the magnetization curve. The relationship between the applied voltage and the flux density depends on the frequency and core cross section according to the well-known equation

$$\left.\begin{matrix}\text{Effective value of}\\ \text{applied voltage}\end{matrix}\right\} = 4.44fNBA \times 10^{-8} \tag{160}$$

where

N = number of turns

f = frequency

B = flux density in the core, in lines per unit area

A = net area of the core.

It is apparent that the voltage which can be applied for a given permissible flux density is proportional to the frequency, so that the voltage limit occurs at the lowest frequencies to be amplified. The permissible flux density depends upon the amount of distortion that can be tolerated, the characteristics of the iron, the length of the air gap (large air gap reduces non-linearity of the inductance and permits higher flux densities), and the circuits associated with the transformer. It is commonly found that, with commercial output transformers operated within their power rating, appreciable distortion will occur at frequencies that the transformer amplifies satisfactorily.

58. Class A Power Amplifiers Using Pentode and Beam Tubes.— Class A power amplifiers using pentode tubes require special consideration because the plate current of a pentode is substantially independent of plate voltage except at low plate voltages. As a consequence, the dynamic characteristic is as shown in Fig. 162, and follows the characteristic curves of the tube irrespective of the load resistance except when the plate voltage becomes low. As compared with the corresponding characteristic of a triode shown in Fig. 152, the pentode dynamic characteristic has much greater curvature, and, instead of becoming straighter as the load resistance is made high, develops an inflection point and becomes more curved.

The plate current of a pentode at the operating point should be as large as the tube rating will permit, and the screen voltage should be such

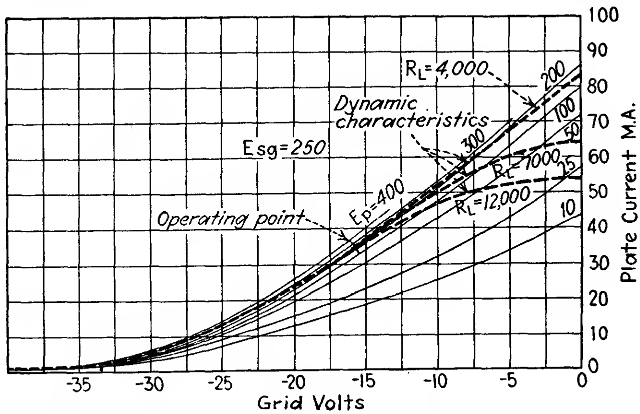


FIG. 162.—Dynamic characteristic of pentode power amplifier for several values of load resistance.

that, when the grid potential has an instantaneous value corresponding to the negative crest of the signal, the plate current will have the minimum allowable value. The usual tube is designed so that the proper screen potential equals the rated plate voltage. The proper grid bias is midway between the grid bias at which the plate current becomes so small that the dynamic characteristic has excessive curvature and the grid potential (either zero or positive as the case may be) at which it is desired to terminate the other end of the portion of the dynamic characteristic that is utilized. The difference $I_0 - I_{\min}$ between the plate current at the operating point and the minimum allowable plate current represents the crest value of the alternating current flowing through the load impedance when the signal voltage equals the grid bias. The load resistance should be such that with this alternating current the minimum potential $E_B - R_L(I_0 - I_{\min})$ reached by the plate is the value at which excessive reverse curvature begins (see Fig. 162). This gives

$$\left.\begin{array}{l}\text{Proper load resistance for pen-}\\ \text{tode power amplifier}\end{array}\right\} = R_L = \frac{E_B - E_{\min}}{I_0 - I_{\min}} \tag{161}$$

where $E_{\min}$ is the minimum allowable plate potential when the instantaneous grid potential is zero if the grid is not to go positive, or at the appropriate positive grid voltage if grid current is to be drawn. The power output is the square of the effective current times the load resistance, and so is approximately,

$$\left.\begin{array}{l}\text{Power output of}\\ \text{power pentode}\end{array}\right\} = \frac{(I_0 - I_{\min})^2 R_L}{2} = \frac{(I_0 - I_{\min})(E_B - E_{\min})}{2}$$
$$= \frac{(E'_{\max} - E_{\min})(I_{\max} - I_{\min})}{8} \tag{162}$$

The input power is $E_B I_0$, so that

$$\left.\begin{array}{l}\text{Plate efficiency of}\\ \text{power pentode}\end{array}\right\} = \frac{1}{2}\left(1 - \frac{I_{\min}}{I_0}\right)\left(1 - \frac{E_{\min}}{E_B}\right) \tag{163}$$

These equations are analogous to Eqs. (146), (149), (154), and (155), and apply irrespective of whether or not there is grid current. If the screen-grid current is taken into account, the effective efficiency is somewhat less, perhaps 80 per cent of that calculated from Eq. (163).

In practical tubes, $I_{\min}/I_0$ and $E_{\min}/E_B$ have values between 0.15 and 0.20 under proper operating conditions. The proper load resistance is hence approximately E_B/I_0 (*i.e.*, the ratio of d-c plate voltage to d-c plate current), and is much less than the plate resistance, while the plate efficiency is about 35 per cent when the screen current is neglected and about 30 per cent when the screen current is taken into account. This relatively high plate efficiency results from the fact that the plate can draw a large current when at a relatively low potential because the electrons are drawn from the space charge by the screen grid, and the plate need be only sufficiently positive to collect all the electrons that approach it.

Pentode power tubes ordinarily require a smaller driving voltage than triode power tubes. This is because in triodes the amplification factor must be low to enable the plate to draw a large current when at a low potential, whereas in the pentode the amplification factor μ_{sg} can be higher because the screen grid is always at a relatively high voltage.

The distortion in a pentode power amplifier and also the exact value of the undistorted part of the power output can be accurately determined for resistance loads by using Eq. (152) in conjunction with a load line drawn upon the E_p-I_p characteristic curves of the tube as shown in Fig. 163. These lines pass through the operating point and intersect the zero-plate-

voltage axis at a current of $I_0 + E_B/R_L$ and the zero-current axis at a voltage of $E_B + I_0R_L$, where I_0 is the current at the operating point, E_B is the plate-supply voltage, and R_L is the load resistance.

The price paid for the greater efficiency and sensitivity of the pentode is greatly increased distortion, which is much worse than with triodes, as is apparent from the curvature of the dynamic characteristics shown in Fig. 162.[1] The distortion furthermore includes considerable third harmonic, so that a push-pull arrangement will not help much. The distortion is quite sensitive to load impedance, increasing rapidly as the load impedance is increased. As a result, the pentode is used only where

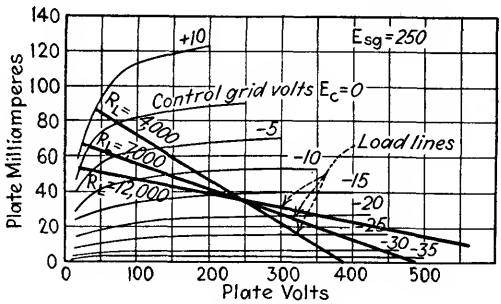


FIG. 163.—Load lines for pentode power amplifier of Fig. 162.

economy is much more important than perfection of performance, or unless negative feedback is employed to reduce the distortion.

Tubes used as pentode power amplifiers must be capable of dissipating appreciable power at the plate, and they are normally designed to operate with a screen-grid potential approximately equal to the plate voltage. The electrostatic shielding between control grid and plate need not be particularly high, and very commonly a certain amount of stray direct coupling is present. The characteristics of typical commercial pentode power-amplifier tubes are given in Table VIII.

Beam Tubes.—The beam tube has characteristic curves similar to those of a pentode. Consequently amplifiers employing beam tubes are designed and analyzed exactly as though pentode tubes were being used, and the same general considerations hold in both cases. The only difference is that in beam tubes there is an abrupt rather than gradual transition between the region where the plate current is substantially independent of plate voltage and the region where the plate current is determined primarily by plate voltage, as discussed in Sec. 33*a*. This reduces the tendency to produce third-harmonic distortion which is

[1] This high distortion of pentodes is further emphasized by experimental results such as those presented by J. M. Glessner, Performance of Output Pentodes, *Proc. I.R.E.*, vol. 19, p. 1391, August, 1931.

characteristic of the pentode, while retaining the advantages of high efficiency and small driving voltage. The second-harmonic distortion in the beam tube is relatively high, but can be eliminated by the use of a push-pull connection. If, in addition, negative feedback is used to reduce the harmonic output, the distortion of a power amplifier employing beam tubes can be made very small.

59. Push-pull Class A and Class AB Amplifiers.—In the push-pull amplifier two tubes are arranged as shown in Fig. 164. The grids are

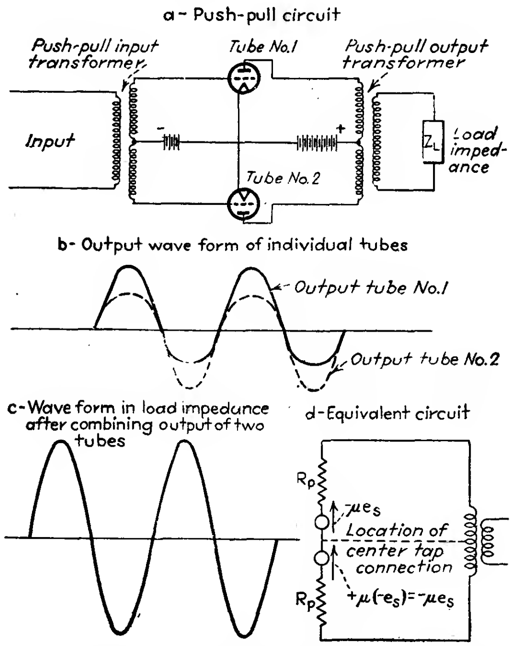


FIG. 164.—Circuit diagram of push-pull amplifier, together with equivalent circuit, and wave shapes produced, showing how the push-pull connection makes the positive and negative halves of the output wave have the same shape even though this is not true of the outputs of the individual tubes. The result is that the output wave contains no even harmonics, and hence suffers less distortion than do the outputs of the individual tubes.

excited with equal voltages 180° out of phase, and the outputs of the two tubes are combined by means of an output transformer having a center tap. The advantages of the push-pull connection, assuming identical tubes, are:

1. No direct-current saturation in the core of the output transformer. (The direct currents in the two halves of the primary magnetize the core in opposite directions, and so produce zero resultant magnetization.)

2. There is no current of signal frequency flowing through the source of plate power. This means the push-pull power amplifier produces no regeneration even when there is a plate impedance common to the power and other stages; it also means that no by-pass condenser is required across cathode-biasing resistor.

3. Alternating-current-hum voltages present in the source of plate power produce no hum in the output because the hum currents flowing in the two halves of the primary balance each other.

4. There is less distortion for the same power output per tube, or more power output per tube for the same distortion, as a result of cancellation of all even harmonics and even-order combination frequencies.

These advantages are so important that a push-pull amplifier using two small tubes is nearly always used in preference to a single larger tube capable of developing the same total power output.

The reason for the elimination of the even harmonics in the push-pull amplifier can be understood by reference to Fig. 164. Assuming that the amplifier is sufficiently overloaded so that some distortion occurs, the individual tubes develop output waves as shown in Fig. 164*b*. The sum of these waves, which represents the amplified output, is shown in Fig. 164*c*; when the two tubes have identical characteristics, the sum will have positive and negative half cycles that differ in sign but not in shape; thus it contains only odd harmonics. This similarity in the shapes of the positive and negative half cycles in the combined output results from the fact that at points one-half cycle apart the tubes have merely interchanged functions, *i.e.*, at the later time tube 2 is operating under exactly the same conditions as was tube 1 a half cycle earlier, and vice versa. A more complete analysis based upon the power-series method shows that not only are the even-order harmonics eliminated, but also all even-order combination frequencies, particularly the sum and difference frequencies. This comes about because second, fourth, and other even-order terms produce current components in the plate circuit having a phase unchanged by reversing the exciting voltage of the tube. The even-order effects are hence of the same phase in both tubes and therefore subtract in the transformer secondary, whereas the odd-order effects add.

Class AB Push-pull Amplifiers.—The elimination of the even harmonics by the push-pull arrangement makes it possible to extend the operating range into the shaded region of Fig. 152, which represents high distortion with single-tube operation. This can be carried to the point where the plate current is actually cut off for a small portion of each cycle without causing excessive distortion. This increases the power output obtainable, particularly when the grid is driven positive, and results in plate efficiencies that range from 40 to 50 per cent. A push-pull amplifier operating in this way is termed a Class *AB* amplifier.

The Class AB amplifier has a large "rectified" plate current when the full signal voltage is applied. This is an advantage since it permits the power input to the plate to be increased in the presence of an applied voltage to a value exceeding the allowable input with no signal. Thus if the "rectified" current equals the no-signal current and the plate efficiency

approaches 50 per cent, the power output obtainable in Class AB operation is approximately equal to the allowable plate dissipation instead of being only 15 to 40 per cent of this as in single-tube operation. The large increase in plate current caused by the signal in a Class AB amplifier, however, makes it impossible to realize the full possibilities of the tube when self-bias is used, since, if the bias keeps the plate current within an allowable value for no applied signal, the bias will be too negative when the "rectified" current flows.

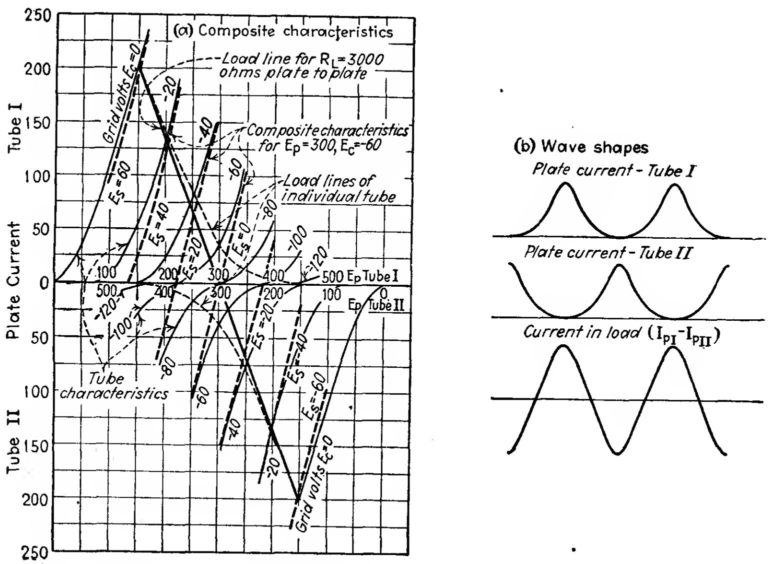


FIG. 165.—Composite characteristic curves of push-pull amplifier with load lines for composite characteristic, together with current waves for the individual tubes.

Analysis of Push-pull Amplifiers and Circuit Arrangements.—The exact analysis of push-pull amplifiers can be carried out with the aid of equivalent circuits and dynamic characteristics. The equivalent circuit of a push-pull amplifier is shown in Fig. 164*d*, and is seen to be equivalent to two tubes in series. The output transformer is hence proportioned to operate out of an effective plate resistance twice that of a single tube, and, if the push-pull tubes are operated under the same conditions as single tubes would be, the proper load impedance from plate to plate terminals is twice that which would be used with a single tube. An accurate analysis for determining best load impedance, proper grid bias, plate efficiency, and distortion requires the use of a composite dynamic characteristic of the push-pull combination, such as shown in Fig. 165.[1]

[1] For further details of this method, see B. J. Thompson, Graphical Determination of Performance of Push-pull Audio Amplifiers, *Proc. I.R.E.*, vol. 21, p. 591, April, 1933.

Here the E_p-I_p curves of the individual tubes are placed back to back, as shown, with the common operating plate voltage superimposed. The composite characteristic curves are then derived by averaging the plate current for grid-potential curves corresponding to the same applied signal voltage, as shown. Thus the curves for the operating grid potential ($E_c = 60$ in Fig. 165) are averaged because they correspond to no signal being applied; but when the signal is 20 volts on one tube, it is -20 volts on the other, and one averages the grid curve for -40 volts on one tube with the curve for -80 volts on the other tube, etc. The resulting derived curves represent the relation between plate-cathode voltage (which is one-half the plate-to-plate voltage) and the difference between the currents to the two anodes (which is twice the a-c current in a plate-to-plate load resistance).

The power output and distortion can be obtained in the usual way from load lines drawn on this diagram. These load lines must pass through the point on the composite characteristic corresponding to the operating plate and grid voltages, and they have a slope corresponding to one-fourth of the plate-to-plate load resistance. To obtain the plate efficiency, it is necessary to know the d-c plate current existing in the presence of a signal. This depends on the shape of the plate-current wave of the individual tube, which can be derived by obtaining the plate voltage for any given instantaneous grid voltage, using the load line drawn on the composite characteristic, and then, from this and the characteristic curve of the individual tubes, finding the resulting plate current. Waves obtained in this way are shown in Fig. 165*b*, while the dotted lines on Fig. 165*a* give the way in which the plate current of the individual tube varies with the instantaneous grid potential during operation (*i.e.*, the dotted lines are in effect load lines for the individual tube).

The usual method of driving a push-pull amplifier is by the use of a center-tapped transformer operating from a triode tube, as shown in Fig. 164. It is possible, however, to avoid the transformer by the use of a "phase reversing" tube to excite the second push-pull tube with a voltage equal in magnitude but opposite in phase to the exciting voltage of the first tube, as shown in Fig. 166. When the grids of a push-pull amplifier are driven positive, it is necessary that the internal impedance of the exciting voltage be low to avoid excessive distortion, exactly as discussed for positive-grid operation of a single-tube amplifier in Sec. 56. It is also necessary to provide a path from grid to cathode having low direct-current resistance so that the grid current will not produce appreciable bias. This means that in circuits such as that of Fig. 166 the grid leaks must be replaced by chokes if the grids are to be driven positive.

The input and output transformers used with push-pull circuits are commonly ordinary interstage and output transformers, respectively,

with one winding center-tapped. The winding that is center tapped should preferably be symmetrical about its mid-point, particularly in the case of the input transformer. This is not usually done in the cheaper transformers such as are used in broadcast receivers, however, and as a result such transformers introduce an unbalance between the tubes at high frequencies, as discussed in Sec. 45. The output transformer has relatively little direct-current saturation, since only the difference between the plate currents of the two tubes produces magnetization in the transformer core. As a result the core of the output transformer can

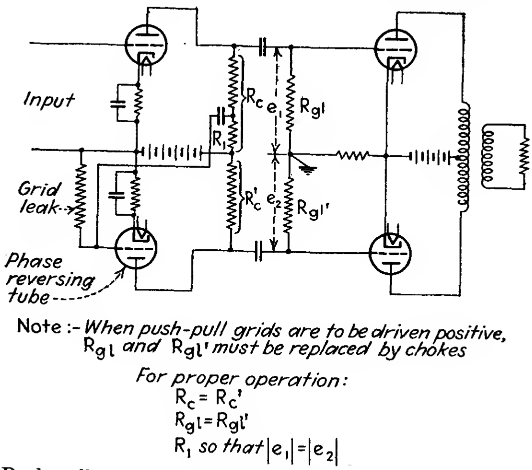


FIG. 166.—Push-pull amplifier excited with the aid of a phase-reversing tube.

be assembled with the smallest feasible air gap, thereby making possible a high primary inductance in proportion to leakage inductance.

Shunt feed is sometimes employed in push-pull amplifiers using large tubes in order to eliminate the d-c current from the primary of the transformer and hence to allow the use of smaller wire for the primary winding than would otherwise be permissible. Shunt feed is also required when the core of the transformer is of high-permeability alloy, because such alloys will have their magnetic properties permanently damaged by severe magnetization such as could occur if one of the tubes accidentally burned out.

60. Class B Audio-frequency Power Amplifiers.[1]—The Class B audio-frequency amplifier is a push-pull amplifier in which the tubes are biased approximately to cutoff. Operated in this manner, one of the tubes

[1] For further information, see Loy E. Barton, High Audio Output from Relatively Small Tubes, *Proc. I.R.E.*, vol. 19, p. 1131, July, 1931; Application of Class B Audio Amplifier to A-C Operated Receivers, *Proc. I.R.E.*, vol. 20, p. 1085, July, 1932; Recent Developments of the Class B Audio- and Radio-frequency Amplifiers, *Proc. I.R.E.*, vol. 24, p. 985, July, 1936.

amplifies the positive half cycles of the signal voltage while the other amplifies the negative half cycles, with the output transformer combining these in the output as shown in Fig. 167. Such an amplifier is characterized by a high plate efficiency and, when the grid is driven positive, as is normally the case, by an unusually high output in proportion to the size of the tube.

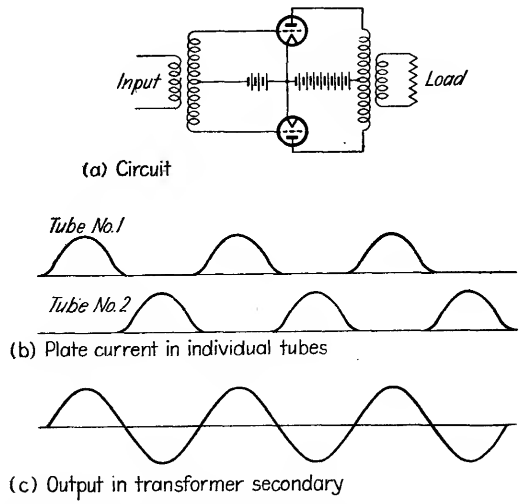


FIG. 167.—Circuit diagram of Class B amplifier, together with oscillograms showing how the half sine-wave pulses of plate current in the individual tubes combine to produce the output.

The output power, plate efficiency, and proper load resistance in a Class B amplifier can be determined with an accuracy sufficient for ordinary purposes by assuming that the characteristic curves of the tube are straight lines as shown in Fig. 168*b*. Analysis based upon this simplifying assumption shows that[1]

[1] These equations are derived as follows: Since the plate-current pulse of each individual tube flows through only one-half of the transformer primary, the combined output of the two tubes is equivalent to an alternating current having a crest value $I_{max}/2$ flowing through the entire transformer primary. If R_L is the equivalent load resistance between primary terminals of the output transformer, the alternating drop produced between plate and cathode of a single tube is one-half the voltage drop of the current $I_{max}/2$ in the resistance R_L or $R_L I_{max}/4$. The minimum instantaneous plate potential is

$$E_{min} = E_B - \frac{R_L I_{max}}{4}$$

Solving this for R_L results in Eq. (164). The power output is one-half the square of the load current times the load resistance, or

$$\text{Power output} = \frac{R_L I_{max}^2}{8}$$

Equation (165) results when R_L is eliminated by the use of Eq. (164). The d-c plate

$$\left.\begin{array}{r}\text{Proper load resistance}\\ \text{from plate to plate}\end{array}\right\} = R_L = 4\frac{E_B - E_{\min}}{I_{\max}} \tag{164}$$

$$\left.\begin{array}{r}\text{Power output from}\\ \text{two tubes}\end{array}\right\} = \frac{I_{\max}(E_B - E_{\min})}{2} \tag{165}$$

$$\text{Plate efficiency} = \frac{\pi}{4}\left(1 - \frac{E_{\min}}{E_B}\right) \tag{166}$$

Here $I_{\max}$ is the peak plate current of the individual tube, $E_{\min}$ is the minimum instantaneous plate potential reached during the cycle, and

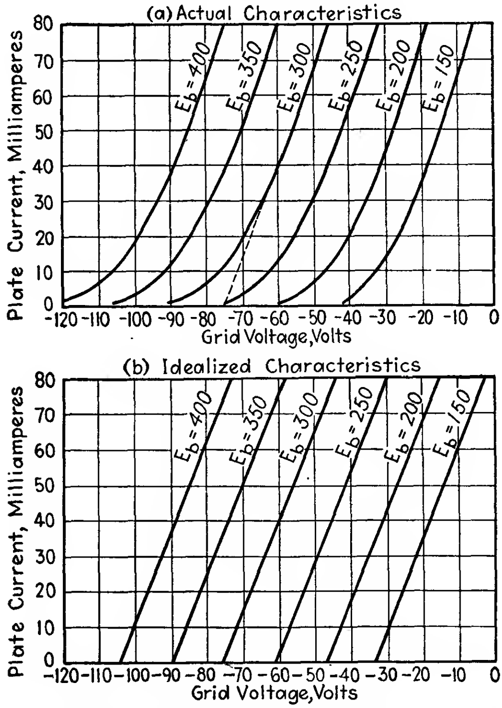


FIG. 168.—Actual and idealized characteristic curves involved in Class B amplifier. In the case of the actual curves the projected cutoff bias is indicated for $E_b = 300$.

E_B is the plate-supply voltage. It will be noted that the maximum possible efficiency is $\pi/4$, or 78.5 per cent, and that the closeness with which the actual efficiency approaches this theoretical maximum is determined by the ratio $E_{\min}/E_B$.

current drawn by the individual tube, assuming a sine-wave half cycle of current, is $I_{\max}/\pi$, so that the total d-c plate current of the two tubes is $2I_{\max}/\pi$ and the power input is $2I_{\max}E_B/\pi$. Dividing the output by this input gives the plate efficiency as in Eq. (166).

The amplitude distortion present in the output of a Class B amplifier depends upon the departure of the tube characteristics from straight lines and can be determined most readily by making use of composite characteristic curves of the push-pull combination derived from the characteristic curves of the individual tubes, as shown in Fig. 165 and discussed above. Study of the distortion based upon composite characteristic curves of the Class B amplifier shows that the best bias to employ is that which corresponds roughly to the cut-off bias that would be obtained if the main part of the characteristic curves were projected as straight lines as in Fig. 168*a*. When this adjustment is employed and when the two tubes have identical characteristics, the output is substantially distortionless as long as the minimum instantaneous plate potential exceeds the maximum positive potential reached by the grid. It is to be noted that, when the bias is the "projected cut-off" value, one has substantially the same conditions existing as were used in deriving Eqs. (164), (165), and (166) except for the fact that the plate efficiency is reduced slightly by the presence of a small plate current when no signal is applied.

The frequency response of a Class B audio amplifier is determined by the characteristics of the output transformer in much the same way as for a Class A power amplifier. There is a falling off at low frequencies as a result of low primary impedance at low frequencies and a falling off at high frequencies as a result of leakage inductance. An exact analysis is involved, however, because the currents flowing through the two halves of the primary are not continuous.[1]

Tubes used for Class B amplifiers operating at plate voltages of about 400 or less are usually designed with a high amplification factor such that "projected cut-off" is approximated by zero bias. In some cases dual-grid tubes with characteristics as in Fig. 169 and discussed in Sec. 36 are employed, while in other cases "twin triodes," consisting of two high-mu triode power tubes in a single envelope, are used. In some of these twin-triode tubes the amplification factor is not particularly high, so that the zero bias current is large and the amplifier operation approaches Class AB rather than true Class B operation. When very large output powers are required, as in the case of high-level modulation of radio transmitters, the same types of tubes are employed for Class B audio-frequency amplification as for Class C radio-frequency amplifiers, and a suitable negative bias is required.

[1] For a discussion of factors determining the amplitude and frequency distortion in a Class B amplifier, see True McLean, An Analysis of Distortion in Class B Audio Amplifiers, *Proc. I.R.E.*, vol. 24, p. 487, March, 1936; A.P.T. Sah, Quasi Transients in Class B Audio-frequency Push-pull Amplifiers, *Proc. I.R.E.*, vol. 24, p. 1522, November, 1936.

The plate current of a Class B amplifier varies with the signal, so that for best results the plate-supply system should have good voltage regulation. This is particularly true when the operating point is not zero grid bias, since then variations in the supply voltage seriously alter the effective operating point.[1]

The driving stage for a Class B amplifier must be designed to have a low internal impedance in order to minimize the distortion resulting from the grid current, as discussed in Sec. 56. The usual driving arrangement consists of a triode power amplifier having an output transformer with a low turn ratio, commonly a step-down. The driver stage of Class B amplifiers employing dual-grid tubes commonly employs a similar dual-grid tube connected as a triode.

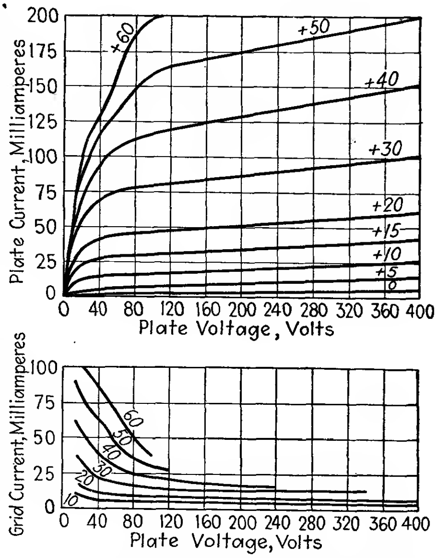


FIG. 169.—Characteristic curves of Type 46 dual-grid tube connected for Class B operation. The figures on curves are grid voltages.

The fundamental operating parameters of a Class B amplifier are the maximum grid potential E_{max} and the minimum plate potential E_{min} reached during the cycle. These two voltages acting simultaneously upon the tube electrodes produce a peak plate current I_{max} determined by the tube characteristics. The proper load resistance is then given by Eq. (164). If the load resistance is higher than this, E_{min} will be lower than assumed, and vice versa. The power output and plate efficiency can be determined from Eqs. (165) and (166), respectively, for known values of I_{max} and E_{min}. The maximum possible power output without excessive distortion and also the maximum possible plate efficiency occur when the maximum grid and minimum plate potentials are approximately equal, but this condition causes the grid current to be large and hence calls for a correspondingly high driving power. For most practical requirements it is more satisfactory to operate with a minimum plate potential E_{min} considerably greater than the maximum positive grid potential E_{max}, since this results in an appreciable reduction in grid current, as is apparent from Fig. 169, and

[1] A method of compensating for the regulation in plate and grid voltages of large Class B amplifiers is described by R. J. Rockwell and G. F. Platts, Automatic Compensation for Class B Bias and Plate Voltage Regulation, *Proc. I.R.E.*, vol. 24, p. 553, April, 1936.

reduces the driving power correspondingly. In actual operation with practical Class B amplifiers, it is necessary to compromise between such factors as freedom from distortion, amount of driving power required, power output, efficiency, etc. The best compromises for smaller tubes are tabulated by the manufacturers of tubes, but must be worked out in each individual case with large tubes.[1]

Compared with Class A power amplifiers, the Class B arrangement has the advantage of higher plate efficiency and negligible power loss when no signal voltage is applied, but possesses the disadvantages that, in order to insure low distortion, it is absolutely necessary that the desired operating conditions be very closely realized. In particular, the two tubes must be very closely balanced, since a difference of 10 per cent in the plate currents will produce 5 per cent second-harmonic distortion; it is also very necessary that the load resistance have the proper value for all frequencies and that the exciting voltage applied to the grids have the proper amplitude. As a consequence Class B audio amplifiers are extensively used in radio transmitters, where the equipment is operated under the continuous supervision of trained personnel, but they are not favored for radio receivers, small public address systems, etc.

61. Class C Tuned Amplifiers.—The Class C tuned amplifier differs from an ordinary tuned amplifier in that the bias is made greater than the cutoff value corresponding to the plate-supply voltage, so that, when a signal is applied, the plate current flows in pulses that last for less than half a cycle. The Class C amplifier is characterized by high plate efficiency and is used to develop radio-frequency power when a proportionality between input and output voltages is not required.

Voltage, Current, and Impedance Relations in Class C Amplifiers.—The voltage and current relations of a Class C amplifier can be understood by considering oscillograms such as those of Fig. 170. The voltage actually applied to the grid electrode of the tube consists of the grid bias E_c plus the exciting voltage E_s. The relations are normally such that at the crest of the cycle the grid is driven appreciably positive and consequently draws some grid current. The voltage actually appearing at the plate of the tube consists of the battery voltage E_b minus the voltage drop E_L in the plate load impedance; thus it has the wave shape shown in Fig. 170*b*. The phase relations are such that the minimum instantaneous plate potential E_{min} occurs at the same part of the cycle as the maximum grid potential E_{max}. The alternating components of the plate and grid voltages are also always sinusoidal since they are developed across sharply resonant circuits.

[1] Such analyses for typical water-cooled tubes are given by I. E. Mouromtseff and H. N. Kozanowski, Comparative Analysis of Water-cooled Tubes as Class B Audio Amplifiers, *Proc. I.R.E.*, vol. 23, p. 1224, October, 1935.

The plate and grid currents that flow at any instant are the result of the combined action of the plate and grid potentials at that instant, and they can be determined from these potentials with the aid of a set of complete characteristic curves of the tube. The plate current is in the form of an impulse flowing for something less than half a cycle. The grid current flows only when the grid is positive, and it is usually sharply

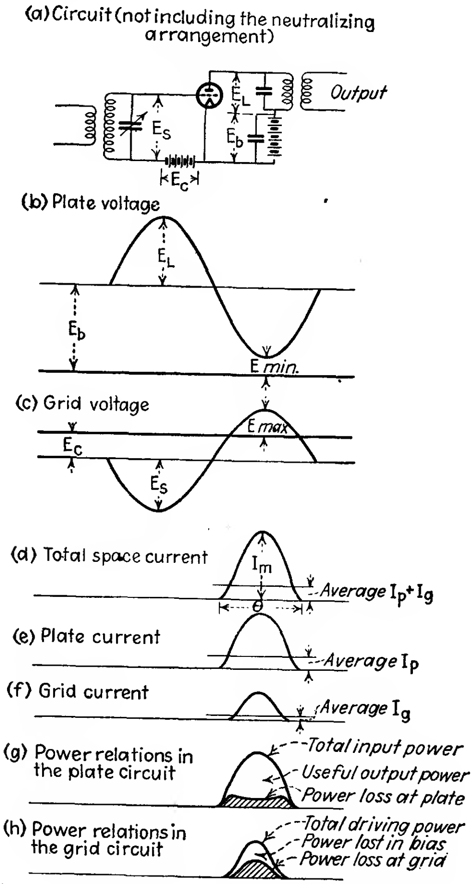


FIG. 170.—Voltage, current, and power relations of typical Class C amplifier.

peaked. In some cases the grid current will reverse and be negative for a portion of the time as a result of secondary emission. The sum $(I_p + I_g)$ of plate and grid currents represents the total space current flowing away from the filament and always has its peak at the instant when the grid and plate potentials are E_{max} and E_{min}, respectively, as shown at Fig. 170*d*. The average value of the plate-current pulse

over a complete cycle represents the direct current that will be drawn from the source of plate power. The average value of the grid-current pulse over a complete cycle is likewise the d-c grid current which will be observed if a direct-current instrument is placed in the grid circuit.

The impedance that the load must supply to obtain proper adjustment is whatever impedance must be placed in series with the plate electrode of the tube to develop the desired alternating plate voltage with the power output represented by the plate-current pulses. The magnitude of this impedance can be controlled by varying the coupling of the load to the plate tuned circuit or by varying the portion of the tuned circuit connected in series with the plate circuit.

Power Relations.—The power delivered to the amplifier by the plate-supply voltage at any instant is the product of instantaneous plate current and the supply voltage, and so varies in the same way as does the instantaneous plate current, as shown in Fig. 170*g*. Part of this input power is delivered to the tuned circuit and represents useful output, while the remainder is dissipated at the plate electrode of the tube. At any instant the division of the total energy between tuned circuit and tube is in proportion to the voltage drops across these parts of the circuit. Thus the plate loss at any instant is equal to the product of instantaneous plate current and instantaneous plate voltage, and so is shown by the shaded area of Fig. 170*g*, while the unshaded part of the total power area represents the energy delivered to the tuned circuit and available for producing useful output. The average input, output, and plate loss are given by averaging the instantaneous values of Fig. 170*g* over a full cycle.

The high efficiency of the Class C amplifier is a result of the fact that the plate current is allowed to flow only when the instantaneous voltage drop across the tube is low, *i.e.*, energy is delivered to the amplifier only when the largest portion of this energy will be absorbed by the tuned circuit. Since the voltage drop across the tube is low only during a small part of the cycle, it is found that the plate efficiency which can be realized is greater the smaller the fraction of the cycle during which the current flows; if the duration of current flow is very small, the efficiency can approach 100 per cent. However, making the duration of current flow very small reduces the input power and consequently the power output, even though the output obtained is developed at a high efficiency. Hence in practical work it is necessary to compromise between high efficiency and high output. Usual conditions correspond to current pulses lasting for 120 to 150°, corresponding to practical efficiencies of the order of 80 to 60 per cent.

The power required to drive the grid positive in a Class C amplifier comes from the exciting voltage. At any moment the driving power is

equal to the product of instantaneous exciting voltage and instantaneous grid current; thus it varies during the cycle as shown in Fig. 170*h*. Part of this driving power represents energy dissipated at the grid electrode in the form of heat, while the remainder represents energy delivered to the bias battery (or, in the case of grid-leak bias, energy dissipated in the grid-leak resistance). Since the grid current flows at or nearly at the crest of the cycle of the exciting voltage, the average driving power over the cycle is very nearly equal to the product of crest exciting voltage and the d-c grid current.[1]

It is to be noted that secondary emission at the grid reduces the driving power by reducing the grid current, but does not reduce the power dissipation at the grid. This is because, when a primary electron collides with the grid and produces a secondary electron that goes to the plate, energy is dissipated at the grid even though there is no effect on the grid current or grid driving power. Hence in the presence of secondary emission the total grid loss is greater than that calculated on the basis of grid voltage and grid current.

Fundamental Factors Involved in the Design and Operation of Class C Amplifiers.—The fundamental factors controlling the behavior of a Class C amplifier are the maximum space current I_m reached during the cycle (or, what is very nearly the same thing, the peak plate current), the minimum instantaneous plate potential E_{min}, the maximum instantaneous grid potential E_{max}, the number of electrical degrees θ during which the plate current flows, and the plate-supply voltage E_b. These are illustrated in Fig. 170. The load impedance is not a fundamental factor since it is dependent upon the above quantities.

The maximum permissible value of peak space current I_m is determined by the electron emission that the filament is capable of producing. With tungsten filaments it is common practice to make I_m very nearly the full emission from the filament. In the case of thoriated-tungsten filaments the deterioration during life is such that factors of safety of three to seven are common, with the exact value depending upon how thoroughly the tube has been evacuated. The characteristics of oxide-coated filaments vary so much that still higher factors of safety must be employed with them.

The values of maximum instantaneous grid potential E_{max} and minimum instantaneous plate potential E_{min} must be such that with these potentials applied to the grid and plate electrodes, respectively, the resulting space current will be the proper value of I_m. Possible combina-

[1] Experimental data indicate that the actual power is usually not over 10 per cent less than the result given by this simple rule. See H. P. Thomas, The Determination of Grid Driving Power in Radio-frequency Power Amplifiers, *Proc. I.R.E.*, vol. 21, p. 1134, August, 1933.

tions of values required to draw any particular peak space current can be determined from complete characteristic curves of the tube.[1]

While there are many combinations of grid and plate voltages that will draw the same space current, it is desired to make E_{min} as low as practicable in order to reduce the plate losses, but at the same time the minimum plate voltage E_{min} must equal or exceed the maximum grid potential E_{max} or the grid electrode will draw an excessive current. Excessive grid current is to be avoided because it reduces the output power and increases the driving power. When low driving power is important, it is desirable to make E_{min} somewhat greater than E_{max}, since in this way the grid current and hence the driving power are considerably reduced, although at the expense of lowered plate efficiency.

With a given peak space current I_m, maximum grid potential E_{max}, and minimum plate potential E_{min}, the angle of current flow θ determines the plate efficiency, the power input, and power output. A large value of θ increases the output power, makes the plate dissipation greater, and lowers the efficiency. This is because the extra input obtained by increasing θ represents current flowing during portions of the cycle when the instantaneous plate voltage is relatively high.

The value of the grid bias for an angle of flow θ is

$$\text{Grid bias} = E_c = \frac{E_b}{\mu} + \left(E_{max} + \frac{E_{min}}{\mu}\right)\frac{\cos(\theta/2)}{1 - \cos(\theta/2)} \tag{167}$$

where E_{max}, E_{min}, and E_b have the same definitions as above, and μ is the amplification factor of the tube.[2] It will be noted from this equation that the grid bias required increases as the angle of flow is reduced and that excessively large bias voltages are required for values of θ much less than 90°.

The signal voltage required to excite a Class C amplifier has a crest value equal to $E_c + E_{max}$, and so depends upon the angle of current flow,

[1] If complete characteristic curves of the tube are not available, it is possible to make an approximate determination of suitable combinations of E_{min} and E_{max} by extrapolating characteristic curves covering only the negative grid region. This is accomplished by plotting $(I_p + I_g)$ as a function of $\left(E_g + \frac{E_p}{\mu}\right)$ upon logarithmic paper as in Fig. 84. A nearly straight line will be obtained as seen from Eqs. (77) and (78), and this can be extrapolated to the desired total space current I_m.

[2] This equation follows from the fact that at the instant the plate current stops flowing the signal and load voltages are $\theta/2$ degrees from their crest values, so that at this instant the effective anode voltage is

$$\frac{E_b - (E_b - E_{min})\cos(\theta/2)}{\mu} - E_c + (E_{max} + E_c)\cos(\theta/2)$$

and this must equal zero.

being greater as θ is decreased. This increase in signal required as the angle of flow is reduced causes the driving power that the exciting voltage must supply to increase as the angle of flow is reduced and to become excessive at angles much less than 90°.

The plate-supply voltage E_b is of considerable importance in determining the output power and plate efficiency, since increasing the plate-supply voltage increases the voltage drop across the load in proportion

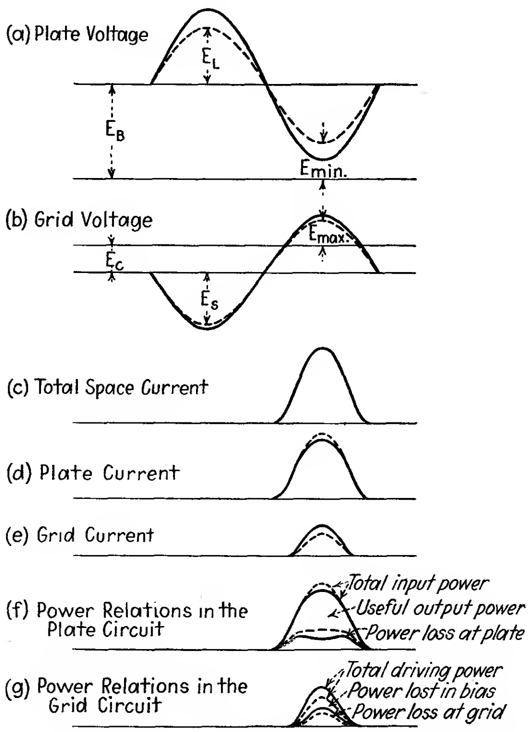


FIG. 171.—Effect of the minimum plate potential in a Class C amplifier. The solid curves are for the same conditions as in Fig. 170.

to the voltage drop across the plate of the tube for the same minimum plate voltage E_{min}. Hence, with given values of θ and E_{min}, increasing the plate-supply voltage will increase the power output greatly while increasing the plate dissipation only slightly. The grid driving power will also be increased somewhat as the plate voltage is made higher, since according to Eq. (167) the grid bias required for a given angle of flow is dependent upon the plate voltage.

Oscillograms showing how the above factors affect the behavior of a Class C amplifier are illustrated in Figs. 171 to 174. Thus Fig. 171 brings out clearly how increasing the minimum plate potential while

keeping the peak space current, angle of flow, etc., the same reduces the grid driving power. Likewise Fig. 172 shows the importance of the angle of current flow in determining the output power, driving power, and plate loss, while Fig. 173 illustrates the great increase in output power with only a small increase in plate dissipation, which results from increasing the plate-supply voltage while keeping the angle of flow, E_{min}, etc., unchanged. Finally Fig. 174 shows what happens when the peak space current I_m is varied by changing the excitation conditions, and it illus-

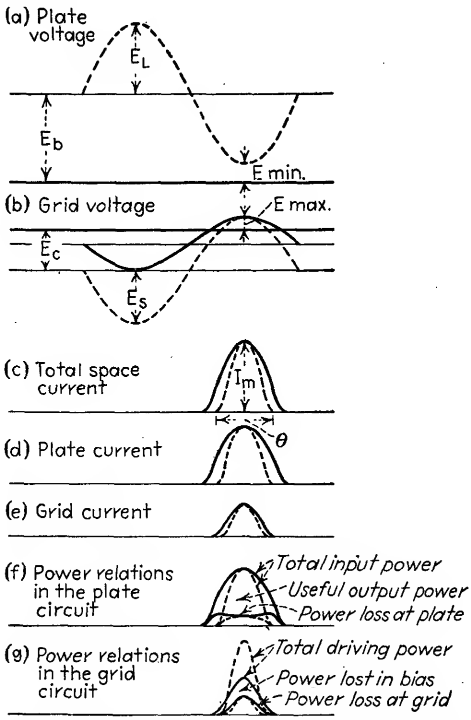


FIG. 172.—Curves showing the effect of the angle of flow in a Class C amplifier. The dotted curves are for the same conditions as Fig. 170.

trates the loss in output with little improvement in efficiency which results when the full emission is not utilized.

Circuit Arrangements.—The plate-tuned circuit, commonly called the "tank" circuit, of a Class C amplifier is usually directly coupled as shown in Fig. 175*a* or is arranged with shunt feed as in Fig. 175*b*. The latter arrangement has the advantage that the coil and condenser are at ground potential for direct-current voltages, but, since the shunt-feed choke is effectively in parallel with the tuning coil, this choke must be a high inductance if it is not to carry an undue proportion of the circulating

current of the resonant circuit. The load to which the output power is delivered may be directly, inductively, or capacitively coupled to the tank circuit (see Fig. 175). Push-pull circuits must be symmetrical on either side of the center tap in order to preserve the balance between the two tubes.

With single tubes, neutralization can be accomplished by one of the systems shown in Fig. 133, while in push-pull amplifiers the cross-neutralization system of Fig. 175*c* is used.

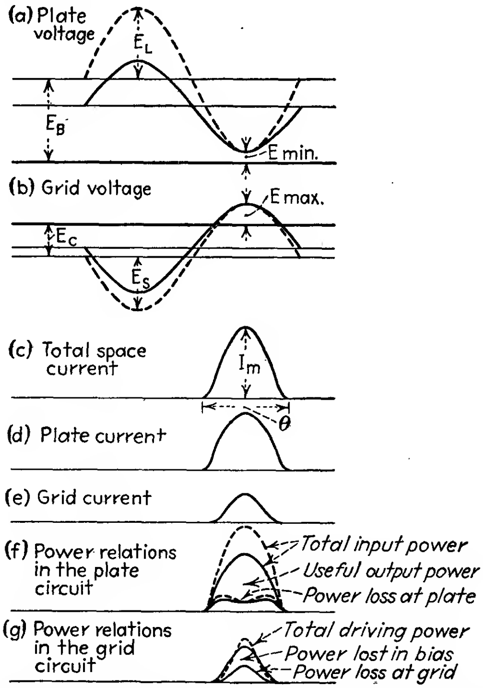


FIG. 173.—Effect of changing the plate-supply voltage in a Class C amplifier. The dotted curves are for the same conditions as in Fig. 170.

The grid bias can be obtained from batteries or a direct-current generator, as illustrated in Fig. 175*a*, or by means of the grid leak and grid condenser, as shown in Fig. 175*b*. In some instances combinations of both methods are used. In the grid-leak method advantage is taken of the fact that the d-c grid current produces a negative bias when passed through a resistance in series with the grid circuit. The magnitude of the bias obtained is equal to the product of d-c grid current and grid-leak resistance, and for a given value of E_{max} it is controlled by the grid-leak resistance. The grid leak must be by-passed to radio-frequency voltages by a condenser appreciably larger than the input capacity of the tube, and also large enough to act as an effective by-pass for the resistance

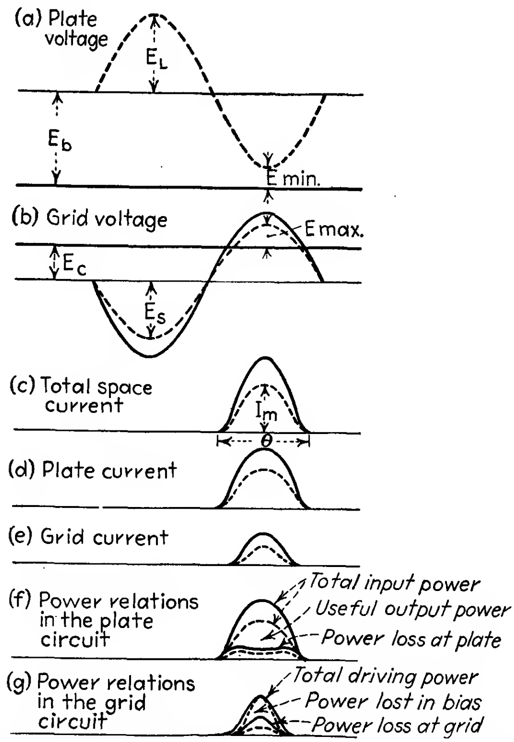


FIG. 174.—Effect of varying the peak space current I_m in a Class C amplifier. The dotted curves are for the same conditions as Fig. 170.

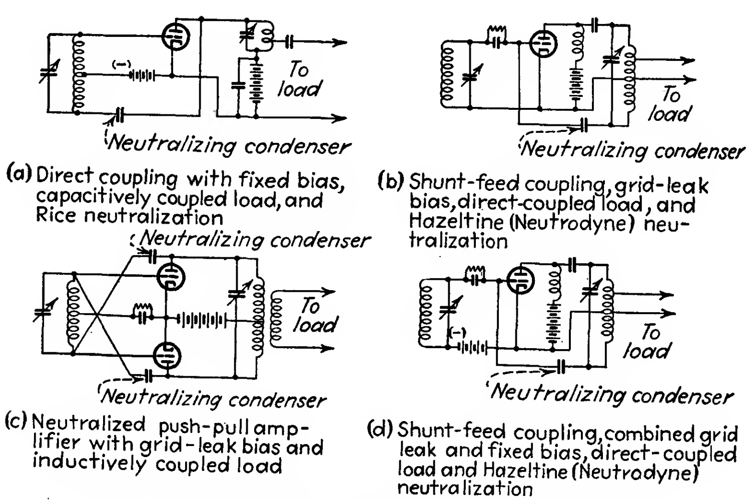


FIG. 175.—Typical circuit arrangements for Class C amplifiers.

used. The grid-leak arrangement has the advantage of simplicity and the fact that it tends to be self-adjusting with respect to maximum grid potential E_{max}. Thus small changes in signal voltage, which would produce large changes in E_{max} with a fixed bias, do not do so with the grid leak because any tendency to change E_{max} produces a large effect on the grid current which tends to change the grid bias in such a way as to maintain E_{max} nearly constant. The disadvantage of the grid-leak arrangement is that, when the exciting voltage is removed, the grid bias is lost. If the resulting high plate current will damage the tube, it is necessary to use either a fixed bias or a combination of grid-leak and fixed bias as shown in Fig. 175*d* or a combination of grid-leak bias and self-bias. When there is a possibility that the grid current will reverse polarity as a result of secondary emission at the grid, as is the case with most water-cooled tubes, grid-leak bias cannot be employed unless a grid-rectifier tube, as illustrated in Fig. 188, is used to prevent the total grid current from becoming negative.

Calculation of Class C Amplifier Performance; Exact Method.[1]—The exact performance of a Class C amplifier, when the plate-supply voltage E_b, minimum plate voltage E_{min}, maximum grid potential E_{max}, and angle of flow θ (or E_c or E_s) are given, can be determined by utilizing complete characteristic curves to derive plate-current, grid-current, plate-loss, etc., curves as illustrated in Fig. 170. The first step in the procedure is to calculate the instantaneous plate and grid potentials at various points of the cycle according to the relations:

$$\text{Instantaneous plate voltage} = e_p = E_b - (E_b - E_{min}) \cos \beta \qquad (168)$$

$$\left.\begin{array}{l}\text{Instantaneous grid}\\ \text{potential}\end{array}\right\} = e_g = (E_c + E_{max}) \cos \beta - E_c \qquad (169)$$

where β is the number of electrical degrees from the crest of the cycle at which e_p and e_g are to be evaluated and where the remaining notation is as before. From these values of instantaneous plate and grid voltages it is possible to plot the corresponding plate and grid currents from complete characteristic curves of the tube, and then to derive the instantaneous plate loss, instantaneous output, instantaneous driving power, etc. Average values can be obtained most readily by plotting curves of the instantaneous values and averaging the results over a full cycle either by using a planimeter or by counting squares to determine the areas involved.[2] The above procedure gives the exact results that

[1] For further information see D. C. Prince, Vacuum Tubes as Power Oscillators, *Proc. I.R.E.*, vol. 11, pp. 275, 405, 527, June, August, and October, 1923.

[2] An alternative procedure for carrying out the averaging process mathematically without drawing the curves is described by D. C. Prince, *loc. cit.*

can be expected, but it has the disadvantage of being laborious and of requiring a complete set of characteristic curves.

Calculation of Class C Amplifier Performance; Approximate Method.[1]—An approximate calculation of Class C amplifier performance can be obtained without point-by-point calculations by taking advantage of the fact that the total space current ($I_p + I_g$) can be expressed rather accurately by Eq. (77), which is repeated below.

$$\text{Total space current} = I_p + I_g = k\left(\frac{e_p}{\mu} + e_g\right)^\alpha \qquad (170)$$

Here e_p and e_g are the instantaneous plate and grid potentials respectively, μ is the amplification factor of the tube (assumed constant), k is a constant, and α is another constant normally having a value very close to $3/2$. To the extent that Eq. (170) holds, the pulses of plate current in a Class C amplifier have a definite form in which the direct-current and fundamental-frequency components are functions only of the angle of flow θ, the peak value I_m of the space current, and the exponent α. This relationship has been worked out and is presented in Fig. 176 for values of α between 1 and 2.[2]

The direct-current component of the total space current divides between the grid and plate electrodes; therefore, to determine the d-c plate current, it is necessary to estimate the d-c grid current. When $E_{\max} = E_{\min}$, the d-c grid current may range up to about 20 per cent of the total space current, but will be less if there is much secondary electron emission at the grid or if the minimum plate potential exceeds the maximum grid voltage.

The fundamental alternating-current component of the total space current is likewise divided between grid and plate electrodes, with the amount going to the grid very nearly equal to twice the d-c grid current. This comes about because most of the grid current flows during the very crest of the cycle, and it can be shown that this represents an alternating

[1] This follows the methods outlined in the following papers: F. E. Terman and Wilber C. Roake, Calculation and Design of Class C Amplifiers, *Proc. I.R.E.*, vol. 24, p. 620, April, 1936; F. E. Terman and J. H. Ferns, The Calculation of Class C Amplifier and Harmonic Generator Performance of Screen Grid and Similar Tubes, *Proc. I.R.E.*, vol. 22, p. 359, March, 1934.

For other methods of approximate analysis, see W. L. Everitt, Optimum Operating Condition for Class C Amplifiers, *Proc. I.R.E.*, vol. 22, p. 152, February, 1934; Burton F. Miller, Analysis of Class B and Class C Amplifiers, *Proc. I.R.E.*, vol. 23, p. 496, May, 1935; A. P. T. Sah, The Performance Characteristics of Linear Triode Amplifiers, *Science Repts. Nat. Tsing Hua Univ., Peiping, China*, vol. 2, pp. 49, 83, April and July, 1933.

[2] The method of deriving this curve, including analysis of the pulses for fractional values of α, is given by Terman and Roake, *loc. cit.*

component that is twice the direct-current value of the grid current.[1] The alternating component of the plate current is hence the alternating-current component of the total space current as obtained from Fig. 176 minus twice the d-c grid current. The power input to the Class C amplifier is now the product of battery voltage and d-c plate current, or

$$\text{Power input} = E_b \times I_{dc} \tag{171a}$$

where I_{dc} is the d-c plate current. Likewise the power delivered to the

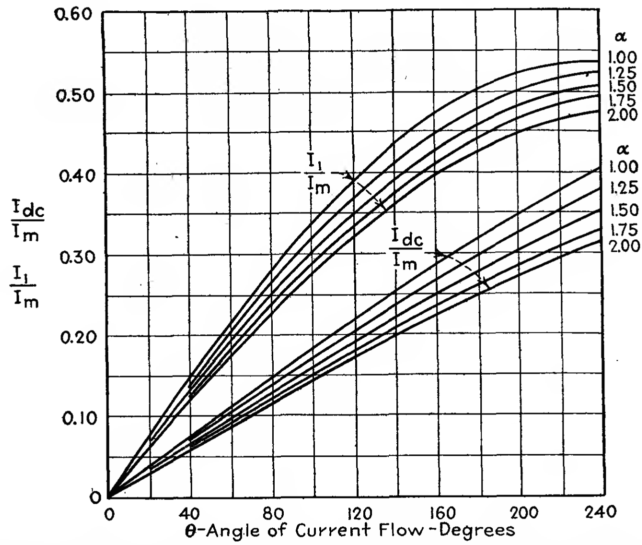


FIG. 176.—Curves giving relation of direct-current and fundamental-frequency component of the space-current pulse as a function of angle of flow θ, and the peak amplitude I_m.

load is equal to half the product of a-c plate current and alternating-current voltage developed across the load, or

$$\text{Power output} = \frac{(E_b - E_{\min})I_1}{2} \tag{171b}$$

where I_1 is the crest value of the fundamental-frequency component of the plate current. The plate dissipation is the difference between these two powers, and the efficiency is their ratio.

The required load impedance is

$$\left.\begin{array}{l}\text{Load impedance between}\\ \text{plate and cathode}\end{array}\right\} = \frac{E_b - E_{\min}}{I_1} \tag{172a}$$

The grid driving power is then approximately equal to the d-c grid current as estimated above times the crest value of the exciting voltage, as previously explained.

[1] See F. E. Terman and J. H. Ferns, *loc. cit.*

The only approximations involved in the method of analyzing Class C amplifiers outlined above are the uncertainty regarding the exact amount of grid current and the assumption that the exponent α in Eq. (170) is constant. The necessity of making a guess as to the grid current does not introduce appreciable error, however, since the grid current is always a small proportion of the total space current. Also, in practice α is found to be very close to $3/2$ over the essential part of the tube characteristic, provided saturation by insufficient emission is not approached. The only practical circumstance where such saturation is found is with tungsten-filament tubes operated so the total space currents are very close to the peak emission available. Even then the error that results is not particularly great, and it alters both power input and power output in about the same proportion, so that the predicted plate efficiency will still be almost exactly correct.

Design of the "Tank" Circuit.—The tuned circuit connected between the cathode and plate of the Class C amplifier, commonly called the *tank circuit,* must supply the proper impedance and must not consume an undue proportion of a power output of the amplifier. The efficiency of the tank circuit is the proportion of the total power delivered to this circuit by the tube that is transferred to the load, and it depends upon the actual loss resistance of the tuned circuit compared with the resistance that is coupled into the tuned circuit by the load. Thus, if the tuned circuit in the absence of load has $Q = 100$, but has $Q = 10$ in the presence of the load, the efficiency of the tank circuit is $(100 - 10)/100 = 0.90$. From the point of view of efficiency, it is desirable that the effective Q of the tank circuit, taking into account the load, be as low as possible, with the actual Q of the circuit, in the absence of load, being as high as is practicable. However, there is a limit to the extent that one can reduce the effective Q, since the lower the circuit Q, the larger will be the harmonic voltages developed across the tank circuit and hence the greater the harmonic energy delivered to the load. Also, if the effective Q is too low, it will be found that the tuning capacity which gives maximum voltage across the tank circuit does not cause the circuit to present a unity-power-factor impedance to the tube. This effect is partly a result of the harmonic voltages developed across the tank circuit and is partly due to the behavior of parallel resonant circuits having low Q, as discussed in Sec. 14. Practical experience indicates that, if the effective Q of the tank circuit exceeds 10 to 12, the maximum impedance will be obtained almost simultaneously with unity power factor.

With the proper tank circuit Q determined by the above considerations, it is then possible to calculate the required value of ωL, knowing the necessary load impedance, and from this to determine the inductance and capacity required. From these and a knowledge of the power

output, the peak voltage developed across the tuning condenser can be calculated by the relation

$$\left.\begin{array}{l}\text{Effective voltage across}\\ \text{tank circuit}\end{array}\right\} = \sqrt{P\omega L Q_{\text{eff}}} \tag{172b}$$

where

P = power in watts delivered to tank circuit
ωL = reactance of inductive branch of tank circuit
Q_{eff} = effective Q of tank circuit, taking into account the effect of the coupled load resistance.

Design and Adjustment.—The design of a Class C amplifier can be systematically worked out on paper according to the following steps:

First.—Select the peak space current I_m on the basis of the electron emission of which the filament is capable, using a suitable factor of safety as discussed above.

Second.—Select a suitable combination of maximum grid potential E_{max} and minimum plate potential E_{min} which will draw this total space current. This is preferably done with the aid of complete characteristic curves of the tube, but it can, when necessary, be carried out by extrapolation, as explained in the footnote on page 318. The minimum plate voltage must not be less than the maximum grid potential, and, if low driving power is important, the minimum plate potential should be appreciably larger than the maximum grid potential, although still relatively small compared with the plate-supply voltage.

Third.—Decide upon a suitable angle of flow θ, making a reasonable compromise between the high efficiency, small output, and large driving power obtained with small angles of flow and the large output, small driving power, and low efficiency with large angles. Under most circumstances the angle of flow will lie between 120 and 150°.

Fourth.—Calculate the grid bias by the use of Eq. (167), which also determines the signal voltage required since the crest signal voltage is $E_c + E_{\text{max}}$.

Fifth.—Determine the d-c plate current, d-c grid current, plate dissipation, power output, efficiency, grid driving power, etc., either by the exact or by the approximate methods.

Sixth.—Examine the results obtained to see if they are satisfactory. If not, revise the design and recalculate the performance.

Seventh.—Estimate the proper Q for the tank circuit, keeping in mind that the effective Q must be at least 10 and that higher values will reduce the harmonic energy delivered to the load but will cause a larger proportion of the output to be dissipated in the tank circuit. With the Q fixed in this way, it is possible to calculate the inductance and capacity required to tune to the desired frequency and develop the required impedance.

The design procedure and the details involved in calculating the performance by the approximate method can be understood with the aid of the following example:

Example.—A Class C amplifier is to be designed employing a Type 800 tube operating at 1000 volts plate potential and having characteristic curves as given in Fig. 75.

The peak emission I_m will be taken as 407 ma, which is arrived at by assuming an electron emission of 100 ma per watt of filament power and a factor of safety of 6. A suitable combination of E_{max} and E_{min} for drawing this current without excessive driving power is $E_{min} = 150$, $E_{max} = 120$. With I_m, E_{max}, and E_{min} determined, the next step is the selection of a suitable angle of flow, which will be taken as 140° as a reasonable compromise between efficiency and output. On the assumption that α equals $3/2$, Fig. 176 gives the direct-current and fundamental-frequency components of the total space current as 0.22 and 0.39, respectively, of the peak space current I_m. The direct-current component of the total space current is hence $407 \times 0.22 = 89.5$ ma, and the crest fundamental-frequency component of the total space current is $407 \times 0.39 = 159$ ma, crest value. It is now necessary to make allowance for the part of the total space current diverted to the grid. Assuming that the d-c grid current is 15 per cent of the total d-c current, the d-c grid current will be 13.5 ma. The d-c plate current will hence be $89.5 - 13.5 = 76.0$ ma, and the fundamental-frequency component of the plate current is, similarly, $159 - 2 \times 13.5 = 132$ ma, crest value. The power input to the plate circuit is the product of the d-c plate current and the plate-supply voltage, or $1000 \times 0.076 = 76$ watts, while the power output is half the product of crest a-c plate current and crest alternating voltage across the load, and so is $0.132(1000 - 150)/2 = 56.1$ watts. The plate loss is $76 - 56.1 = 19.9$ watts, and the efficiency is $56/76 = 73.7$ per cent. The grid bias required as calculated by Eq. (167) is found to be 134 volts. The crest alternating-current driving voltage is $(E_c + E_{max})$ or 254 volts, and the grid driving power is to a first approximation $254 \times 0.0135 = 3.4$ watts. The load impedance that is required is the ratio of alternating-current voltage $(E_b - E_{min})$ to the alternating-current component of the plate current, and so is $(1000 - 150)/0.132 = 6430$ ohms.

The accuracy of the above approximate analysis is indicated by the following comparison with an exact point-by-point analysis:

	By approximate analysis using Fig. 176	By exact point-by-point calculation
D-c plate current	76 ma	74.3 ma
D-c grid current	13.5 ma	13.4 ma
A-c plate current	132 ma	131 ma
Power input	76.0 watts	74.3 watts
Power output	56.1 watts	55.8 watts
Plate loss	19.9 watts	18.6 watts
Plate efficiency	73.7 per cent	75.2 per cent

If the tank circuit is directly coupled as in Fig. 175a and if an effective Q of 12 is chosen, $\omega LQ = Z_L$ and $\omega L = 6430/12 = 536$ ohms. The

corresponding inductance and capacity can then be calculated for any frequency.

Practical Adjustment of Class C Amplifiers.—The adjustment of Class C amplifiers to realize the design conditions is usually carried out by a cut-and-try process using the d-c grid current, the d-c plate current, and the output power as guides.

The procedure to be followed depends somewhat upon whether the bias is developed by a grid leak or by a fixed source. With a fixed bias, the adjustment after setting the bias consists in first tuning the tank circuit to resonance by adjusting for minimum plate current with a moderate exciting voltage, and then adjusting the exciting voltage and effective load impedance by trial until the expected d-c plate current and output power are obtained without excessive grid current. The procedure can be simplified by keeping in mind that the total space current $I_p + I_g$ depends primarily on the maximum grid potential E_{max} and hence upon the excitation, while with a given excitation the minimum plate potential depends upon the load impedance, and hence the ratio of grid current to plate current is less the higher the impedance in the plate circuit of the tube. The adjustment procedure hence involves coupling the load to the tank circuit a reasonable amount, after which the excitation is varied until the total space current approximates the expected value. The effective load impedance is then varied by changing the coupling of the load to the tank circuit until a point is found where decreasing the coupling to the load slightly (*i.e.*, increasing the load impedance) causes the grid current to become excessive. The power output, power input, and grid current are then noted, and the entire procedure is repeated over and over again until the desired operating conditions are obtained. In carrying out the adjustments it is necessary to pay attention to the plate dissipation (*i.e.*, difference between power input and power output), particularly in the early stages, since there is always danger of damaging the tube by accidentally overheating the electrodes.

When the bias is obtained from a grid-leak resistance, added complications result from the fact that the bias and hence the maximum grid potential depend upon the grid current and grid-leak resistance. The best procedure is to start with the grid-leak resistance that is required to produce the desired grid bias with the estimated grid current, and then to follow out the procedure outlined above to obtain the best operating point. The actual grid bias that results is then calculated from a knowledge of the grid-leak resistance and the grid current; if it is not the desired value, the grid-leak resistance is altered as necessary and the process is repeated until the grid bias, space current, power output, and power

input approximate the calculated values and the grid current is not excessive.

The cut and try involved in the above procedures can be practically eliminated by the use of a peak vacuum-tube voltmeter, such as illustrated in Fig. 177, to read directly the amount the grid is driven positive. This makes use of a diode tube, preferably an 879 or similar tube capable of standing a high voltage and having a relatively low electrode capacity. The anode of the diode is connected directly to the grid of the Class C amplifier, while the cathode is biased positively with respect to the amplifier cathode by the potentiometer P until the milliammeter M just begins to show current. Under this condition the cathode potential is substantially equal to the most positive potential reached by the rectifier anode, so that the voltmeter V then reads E_{max} directly.[1] With the exciting conditions adjusted to the desired value, it is a relatively simple matter to adjust the load impedance until the plate current, power output, and plate efficiency approximate the expected values.

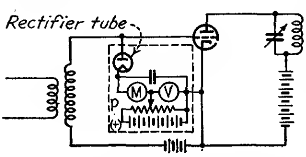


FIG. 177.—Peak voltmeter arranged to determine the maximum positive grid potential E_{max}.

When making adjustments on a Class C amplifier, it is desirable to check the tuning repeatedly, particularly after each change in the load coupling. With triodes this is done by adjusting the tuning condenser for minimum plate current. This condition normally corresponds to maximum load impedance (*i.e.*, to the lowest possible E_{min}), maximum grid current, and a unity-power-factor load. It also gives the maximum possible power output unless the effective Q of the tank circuit is so low that the points of unity power factor and maximum impedance do not coincide.

Neutralization can be accomplished by turning off the filament voltage of the tube (but leaving the tube in place) and applying the exciting voltage. The neutralizing condenser is then adjusted until tuning the tank circuit through resonance has no effect on the d-c plate and grid currents of the exciting stage, or until there is no voltage or current in the tank circuit of the unlighted tube.

Class C Amplifiers Employing Screen-grid Pentode Tubes.—Screen-grid and pentode tubes can be operated as Class C amplifiers by making the grid bias greater than the cut-off value corresponding to the screen-grid potential. The performance obtained is then similar to that of triode Class C amplifiers, but with the advantage that no neutralization is

[1] For further information concerning such peak voltmeters and also for information describing trough meters that will read the minimum plate potential E_{min}, see F. E. Terman, "Measurements in Radio Engineering," 1st ed., pp. 27–29, McGraw-Hill Book Company, Inc.

required. The tubes are expensive in the larger air-cooled sizes, however, and are not available in water-cooled tubes.

The analysis, calculation of performance, and design are the same with screen-grid and pentode tubes as with triodes except for minor modifications introduced by the screen grid. In particular, the relation between the grid bias and angle of flow is

$$\text{Grid bias} = E_c = \frac{\dfrac{E_{sg}}{\mu_{sg}} + E_{\max} \cos\left(\dfrac{\theta}{2}\right)}{1 - \cos(\theta/2)} \tag{173}$$

where E_{sg} is the screen-grid potential and μ_{sg} is as defined by Eq. (66). In selecting operating conditions it is essential that the maximum grid potential $E_{\max}$ should not exceed the screen-grid potential, and with screen-grid tubes the minimum plate potential $E_{\min}$ must also equal or exceed the screen-grid potential. Otherwise the situation is the same as with triodes, and the analysis can be carried out with either point-by-point calculations or the approximate method based upon the curves of Fig. 176.[1]

In adjusting screen-grid and pentode Class C amplifiers, it is sometimes found that the plate potential has so little effect on the plate current that it is impossible to tune the tank circuit by adjusting for minimum plate current. Under such circumstances the tuning adjustment can be made to give maximum current in the load.

62. Characteristics of Tubes Suitable for Use in Class C Amplifiers.—When the amount of power to be generated by a Class C amplifier does not exceed a few watts, it is possible to employ the ordinary small vacuum tubes commonly used in radio receivers; but larger tubes are required for greater powers. The amount of power that a tube can handle is determined by the plate voltage that may be applied to the tube with safety, by the electron emission of the cathode, and by the amount of power that can be dissipated within the tube without overheating.

Filament Size.—The filament must have sufficient electron emission to develop the required value of peak space current I_m throughout the life of the tube. In the case of tungsten filaments, the allowable I_m is taken as being practically the full emission of the filament. With thoriated-tungsten emitters the initial emission is normally between three and seven times the required peak emission to allow for deterioration during the life of the tube. Oxide-coated filaments require still larger factors of safety.

Voltage Requirements.—The voltage that is applied to the plate of a Class C amplifier consists of the direct-current plate-supply potential plus

[1] The details of such an analysis are given by Terman and Ferns, *loc. cit.*

an alternating component that has a crest value nearly equal to the direct-current voltage, so that during operation the instantaneous voltage between plate and cathode varies from nearly zero to nearly twice the plate-supply potential. The voltage that can be applied safely to the plate is limited by the plate-cathode insulation and the degree of vacuum existing within the tube, and, in order to withstand potentials ranging from 1000 to 20,000 volts, it is necessary that extreme care be taken in the design and construction of the tube.

In tubes generating alternating current of very high frequencies, *i.e.*, above 3000 kc, the alternating component of the plate potential produces large dielectric losses in the glass walls because of the very high frequency and the relatively high temperature of the glass during operation of the tube. The result is that, when a tube is operated at a high plate voltage and is generating very high frequencies, there is a tendency for the dielectric losses in the glass to produce local overheating that may soften the glass and destroy the vacuum.[1] Vacuum tubes which are to operate with very high plate voltage must therefore provide ample plate insulation and be arranged to minimize the dielectric stress in the glass walls. Even with the best designs it is usually necessary to use a lower plate voltage when operating at extremely high frequencies than when operating at low or moderate frequencies.

Heat Energy to Be Dissipated.—The plate power supplied to a tube is equal to the product of plate-supply voltage and average plate current and so is limited by the direct-current plate voltage that may be used without danger of a breakdown and by the allowable d-c plate current. A certain fraction of this power supplied to the plate, usually more than half, is delivered to the resonant circuit in the form of alternating-current energy, while the remainder appears at the plate in the form of heat which the plate must be capable of radiating to the walls of the tube without becoming excessively hot. There is also a somewhat smaller power loss at the grid which the grid must be capable of radiating without reaching an excessive temperature. The total power dissipated inside the tube consists of these grid and plate losses plus the power used in heating the cathode, and it must be carried away through the envelope of the tube. Tubes generating 1 kw or less of alternating-current power can transfer the energy loss in the tube to the surrounding air with glass walls of reasonable size, but, since glass softens at relatively low temperatures, the amount of energy that can be radiated per unit area of glass surface is low, and larger tubes would have to be enclosed in glass bulbs of prohibitive size. The cooling of tubes with glass walls is ordinarily obtained

[1] A discussion of punctures resulting from dielectric losses in the glass walls of the tube is to be found in the article by Yujiro Kusunose, Puncture Damage through the Glass Wall of a Transmitting Vacuum Tube, *Proc. I.R.E.*, vol. 15, p. 431, May, 1927.

by allowing free circulation of air, although in the largest sizes a forced draft supplied by a fan is sometimes employed.

A large proportion of the energy dissipated in tubes having glass bulbs is produced at the plate, which must therefore be capable of radiating without damage all the heat generated at its surface plus that fraction of the filament-heating power which the filament radiates to the plate. In order to facilitate this radiation of energy from the plate to the glass walls of the tube, it is desirable to blacken the plate to increase the rate of heat radiation and to use a material that will stand relatively high temperatures. In some types of tubes the plates are actually at a dull red heat when operating under normal conditions.

Water-cooled Tubes.—In tubes having power ratings in excess of 1 kw, the problem of carrying away the energy dissipated in the tube has been solved by using water-cooled plates. Such tubes can be made in several ways. One type of construction is shown in Fig. 178 and employs a cylindrical copper plate which is dropped into a water jacket through which cooling water is circulated. The plate serves as part of the wall of the tube and also acts as an anode; since it is in direct contact with the cooling water, many kilowatts can be dissipated at the plate without an appreciable rise in temperature. Another type of construction that has been commercially successful encloses the tube with a glass bulb but employs a hollow plate through which water is circulated by means of pipes sealed through the glass. The most important feature of these water-cooled tubes is the metal-to-glass seal, which can be made in a number of ways.[1]

While water-cooled plates effectively eliminate the problem of dissipating the energy developed at the plate, there is still the energy dissipated at the grid of the tube, which, while much less than that developed at the plate, is so great in large tubes as to cause the grid to operate at a red heat under normal conditions; it may even heat the grid to a temperature at which thermionic emission of electrons takes place. Tubes with water-cooled grids have been devised and, while not now in commercial use, will probably be employed when ratings in excess of several hundred kilowatts are reached.

Emitters.—The electron emitter used in high-power tubes operating at high plate voltages is practically always a tungsten filament, while oxide-coated and thoriated-tungsten filaments, though having a high thermionic efficiency, are employed only in the smaller tubes. This is because of the effects produced by the gas molecules left in the tube after

[1] There are a number of ways of sealing metal to glass, some of which have been known for many years. For a description of a type of seal that has been extensively used in water-cooled tubes see William G. Housekeeper, The Art of Sealing Base Metals through Glass, *Trans. A.I.E.E.*, vol. 42, p. 870, June, 1923.

evacuation, as explained in Sec. 26. Gas is particularly troublesome in tubes having oxide-coated filaments and makes it impractical to use such emitters when the plate voltage exceeds about 1500 volts.

The amount of power consumed in heating the cathode of a high-power vacuum tube is very large because the required electron emission is great. Data on the cathode-heating power of a number of typical tubes designed for oscillator purposes are to be found in Table IX, which shows that the cathode power is in excess of 1 kw in the largest tubes.

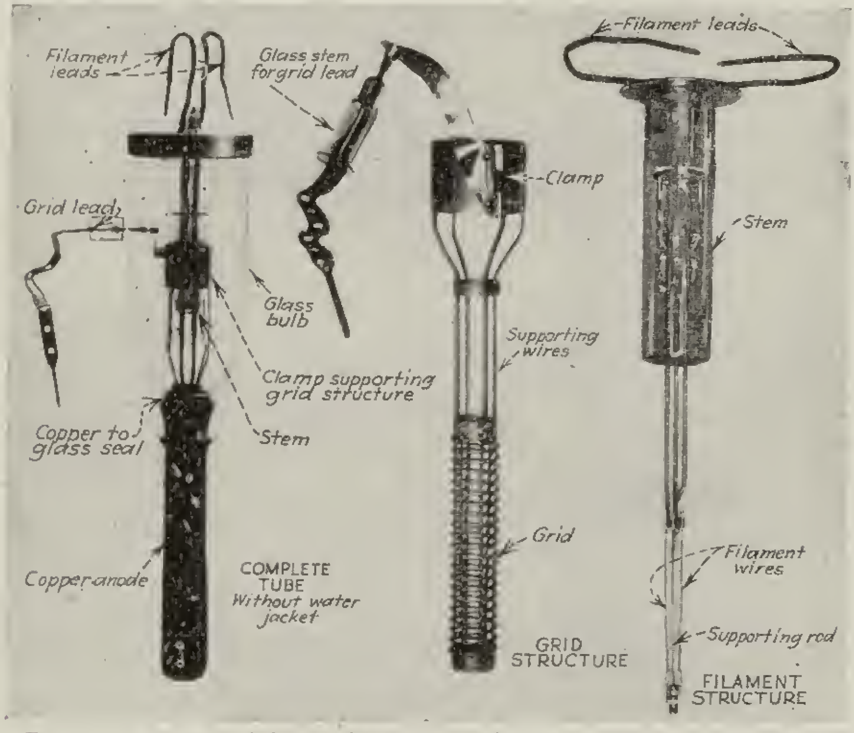


FIG. 178.—A water-cooled type of vacuum tube having a power rating of 20,000 watts output when acting as an oscillator. The anode is made in the form of a copper cylinder which is cooled by immersion in a water jacket.

In large tubes it is customary to place a resistance in series with the filament when the filament circuit is first closed in order to limit the rush of current that would otherwise flow because of the low resistance of the cold filament. Without this starting resistance, the initial current in large tubes would burn out fuses and might even damage the lead wires passing through the glass seal.

Construction and Rating of Tubes.—One of the most difficult problems encountered in the manufacture of high-voltage high-power tubes is the production of the vacuum. The air originally in the tube can be readily

pumped out, but the removal of the gas that is occluded in the metal and glass parts of the tube that are in contact with the vacuum is not a simple matter. Such solid materials will absorb large quantities of gas, which clings very tenaciously and can be removed only by the application of heat. The pumping procedure is carried out with the entire tube in an oven heated to a temperature just below the softening point of the glass in order to remove as much as possible of the gas occluded in the glass parts. Finally, while the tube is still on the pump, the metal parts are brought to temperatures above those that would be produced during normal operation. By taking these precautions it is possible to produce a suitable vacuum with a reasonable certainty that it will be maintained throughout the operating life of the tube. The difficulty of obtaining such a vacuum can be understood when it is realized that the large water-cooled tubes require continuous pumping for approximately 24 hr. to remove the occluded gas.

Materials suitable for the plate and grid electrodes of high-power vacuum tubes are few in number because of the high temperatures to which these electrodes may be subjected during operation. The grid is generally of tungsten because of the refractory nature and high mechanical strength of this material, while molybdenum is also sometimes used. The plates of air-cooled tubes are generally made of molybdenum, although graphite is coming into increasing use. Tantalum is occasionally used for the plates of power tubes and has the desirable property of absorbing gas in a certain range of temperatures. Tantalum plates thus help maintain the vacuum by "cleaning up" any gas that may be given out by the metal and glass parts during operation. The plates of water-cooled tubes do not have to meet any special requirements inasmuch as the operating temperature is low. Copper is generally employed because of its high thermal conductivity, but other materials can be used.[1]

The amplification factors of power tubes do not differ greatly from those of ordinary receiving tubes, although the tendency is to use somewhat higher values because of the much greater plate voltages involved. The plate resistance of large tubes is somewhat lower than for small tubes having similar amplification factors, but the differences are not striking since the increased size of the tube increases the spacing as well as the electrode area. The chief distinguishing features of high-power tubes as compared with small receiving tubes are the very high voltages that the tubes can stand, the large space currents that they are capable of producing, and their ability to dissipate large power losses. The principal characteristics of a representative series of power tubes are given in

[1] For further discussion of anode materials, see E. E. Spitzer, Anode Materials for High-vacuum Tubes, *Elec. Eng.*, vol. 54, p. 1246, November, 1935.

Table IX (Sec. 56), and a number of the same tubes are shown in Figs. 179 and 180. It will be noted that the size of the glass bulb of the air-cooled tubes is proportional to the power rating and that, while the low-power tubes employ a type of construction similar to that used in receiving tubes, the tubes intended for service at high voltages are arranged so

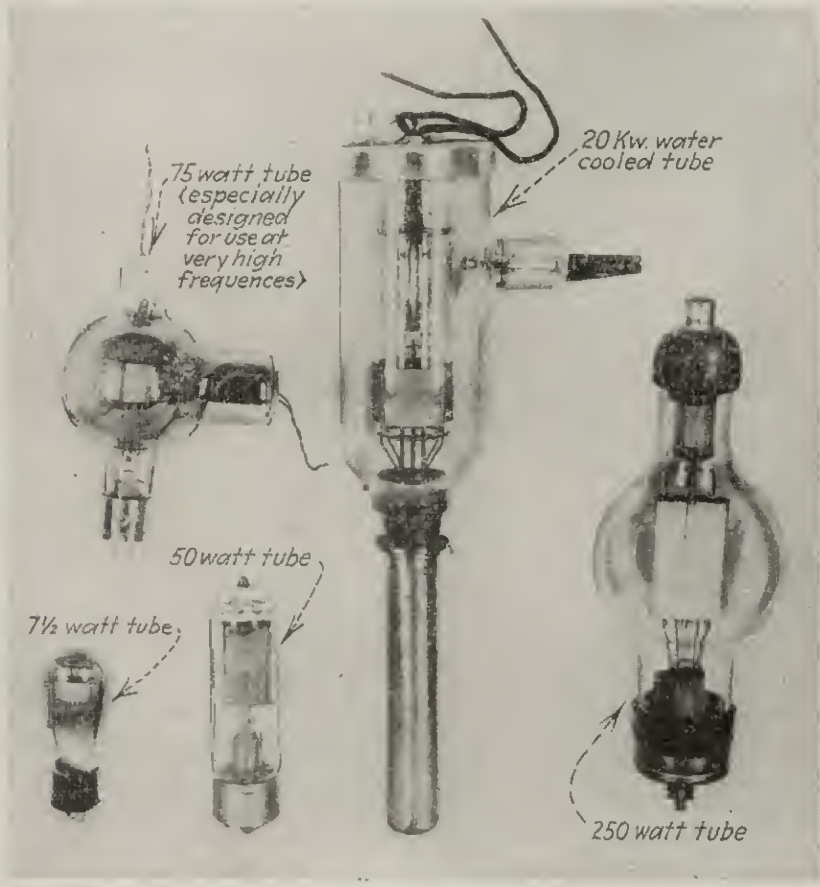


FIG. 179.—Typical oscillator tubes having rated power outputs ranging from 7½ to 20,000 watts. Note that the surface area of the glass bulbs in the air-cooled tubes is proportional to the power rating and that in the larger tubes (*i.e.*, those requiring high plate voltages) the plate leads are brought out through separate seals remote from the grid and filament seals.

that the plate connection enters the tube through a special seal remote from the grid and cathode connections in order to make the insulation strength as great as possible. In many tubes the grid connection is also brought out through a separate seal, which has the advantage of reducing the electrode capacities and increasing the insulation strength still more.

Power tubes are rated on the basis of power output under conservative operating conditions. Thus a 50-watt tube will develop at least 50 watts of radio-frequency output under ordinary conditions. In most cases, particularly with small tubes, ratings are purely conventional, being made on the basis of an anode efficiency of 50 per cent, with a conservative figure taken as the allowable plate loss. Hence a 75-watt tube has a nominal allowable plate loss of 75 watts but can be operated safely with a 100-watt loss, and so at a plate efficiency of 67 per cent it will develop 200 watts of radio-frequency output.

A number of special problems are involved in power tubes intended to be used at very high frequencies. In addition to the necessity of avoiding excessive dielectric loss in the glass walls of the tube, tubes intended for high-frequency service must have low interelectrode capacities in order to allow for external tuning capacity in the resonant circuits. To obtain this low capacity, the electrode areas must be small and the spacing greater than usual, while the different electrodes are preferably mounted from different presses. The 75-watt tube of Fig. 179 is an example of a tube embodying these constructional features. The grid and plate lead-in wires of tubes used at very high frequencies must be capable of carrying heavy currents, because at the frequencies involved the capacity currents flowing to the grid and plate may reach many amperes even though the interelectrode capacities are small.

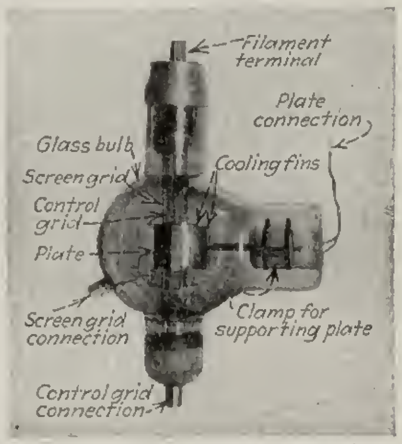


FIG. 180.—Photograph of screen-grid power tube having a rating of 500 watts output when operating as a Class B amplifier.

In power tubes intended to operate at ultra-high frequencies, it is necessary to keep the transit time of the electrons low, as well as to reduce lead inductance and interelectrode capacity. These requirements make it desirable to reduce the physical size, but doing so also reduces the allowable plate dissipation and hence the power rating. The result is that, although considerable progress is being made, the final solution of the ultra-high-frequency power-tube problem does not appear to have been reached.[1]

[1] For an extensive discussion of developments in the field of ultra-high-frequency power tubes, see M. J. Kelly and A. L. Samuel, Vacuum Tubes as High-frequency Oscillators, *Elec. Eng.*, vol. 53, p. 1504, November, 1934.

63. Harmonic Generators.—By taking advantage of the fact that the pulses of plate current have appreciable harmonic content, a class C amplifier can be used to generate output power that is a harmonic of the signal voltage applied to the control grid. It is merely necessary to tune the tank circuit to the desired harmonic and adjust the angle of flow to a value that is favorable for generating the harmonic involved.

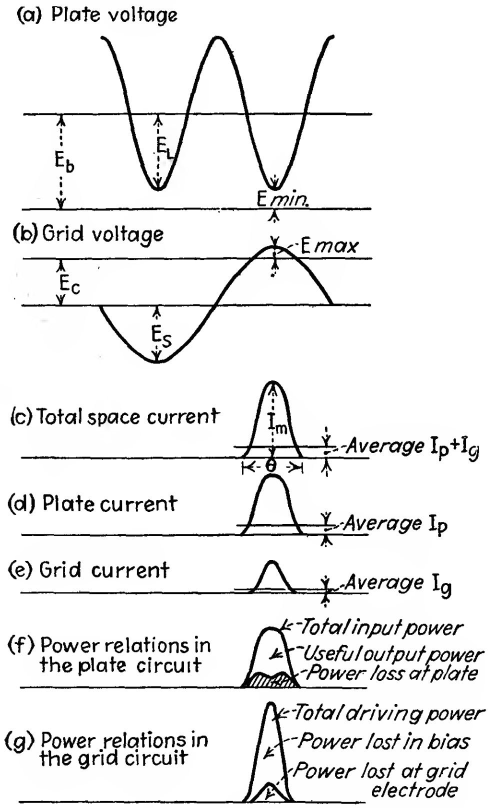


FIG. 181.—Voltage, current, and power relations of typical Class C harmonic generator. Note the similarity to the curves of Fig. 170.

Harmonic generators of this character are frequently used in radio transmitters and for other communication purposes.

Oscillograms showing voltage, current, and power relations in a typical harmonic generator are shown in Fig. 181; they are seen to be almost identical with the corresponding oscillograms of Fig. 170 for a Class C amplifier. The significant factors are still the maximum grid potential E_{max}, the minimum plate potential E_{min}, etc., and the considerations involved in the design and adjustment are essentially the same as in the case of a Class C amplifier except for the fact that, since

the harmonic content of the plate-current pulses depends upon the angle of flow, it is necessary to choose this angle rather carefully in relation to the harmonic to be generated. From the point of view of reasonable output at good efficiency, the best angle of flow is $180/n$ electrical degrees based upon the fundamental frequency, where n is the harmonic to be generated. The output is increased slightly to a maximum when the angle is about 50 per cent greater, or $270/n$ electrical degrees, but this is accompanied by a disproportionate increase in input power and so corresponds to a decidedly lower plate efficiency. The angle of flow can then be increased still further to about $360/n$ electrical degrees before the loss in output becomes excessive, although the efficiency drops rapidly as the angle of flow is increased. It is often desirable to make the angle of flow exceed $180/n$ and even approach $360/n$ in spite of the resulting loss in plate efficiency, because the larger the angle of flow, the smaller will be the grid bias and hence the less the driving power required. The only precaution that must be taken is to make sure that the allowable plate dissipation is not exceeded.

The power output obtainable from a harmonic generator is almost exactly inversely proportional to the order of the harmonic. A properly adjusted second-harmonic generator develops about 60 to 70 per cent as much output as is normally obtained from the same tube operating as a Class C amplifier, so that for third-, fourth-, and fifth-harmonic generation the corresponding percentages are roughly 40, 30, and 25, respectively. This represents appreciable output on high-order harmonics, but, because of the large driving power required when the angle of plate-current flow is small, harmonic generators in commercial radio equipment are usually designed to generate the second harmonic, with the third or fourth harmonic occasionally employed.

Analysis of Harmonic Generator Performance.—The analysis of Class C harmonic generators is carried out in much the same manner as the analysis of Class C amplifiers. The maximum instantaneous grid voltage E_{max}, the minimum instantaneous plate voltage E_{min}, the peak space current I_m, and the angle of flow θ are determined as in the case of Class C amplifiers except for modifications resulting from the considerations outlined above. The bias voltage required for a given set of operating conditions is given by the relation

$$\text{Grid bias} = E_c = \frac{E_b[1 - \cos(n\theta/2)] + E_{min}\cos(n\theta/2)}{\mu[1 - \cos(\theta/2)]} + \frac{E_{max}\cos(\theta/2)}{1 - \cos(\theta/2)} \qquad (174)$$

The notation is the same as in Eq. (167), with the addition that n is the order of harmonic involved. With E_{min}, E_{max}, E_b, θ, grid bias, and

the signal voltage determined, oscillograms of plate and grid currents, instantaneous grid and plate losses, power input, etc., can be drawn from complete characteristic curves of the tube exactly as in the case of a Class C amplifier.

An alternative method of analysis which is relatively simple and gives results sufficiently accurate for design purposes is based on the fact that the *pulse of total space current for a given* $E_{\max}$, $E_{\min}$, E_b, *and* θ *has practically the same shape in the harmonic generator as in the case of a*

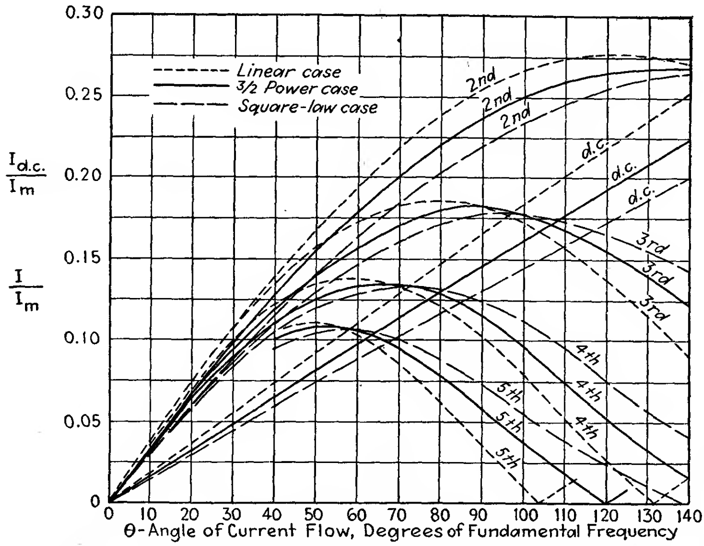


FIG. 182.—Curves giving relation of direct-current and harmonic components of the space-current pulse as a function of angle of flow θ (in electrical degrees of fundamental frequency), and the peak amplitude I_m.

Class C amplifier. The error involved in making this assumption is ordinarily only a few per cent and so can be neglected. The same approximate method of analysis that was used with Class C amplifiers can then be applied to harmonic generators, the only difference being that one is now interested in the harmonic content of the plate-current pulse and total space current instead of the fundamental frequency component. Curves giving the second-, third-, fourth-, and fifth-harmonic amplitudes in the space-current pulse as a function of angle of flow for different values of the exponent α in Eq. (170) are given in Fig. 182. The application of the approximate method of analysis to the case of a Class C amplifier can be understood by the following example:

Example.—A second-harmonic generator will be designed using the same Type 800 tube employed in the Class C amplifier example and operating at a plate potential of 1000 volts.

The same values of I_m, E_{min}, and E_{max} as used in the Class C case can be used, namely, 407 ma, 150 volts, and 120 volts, respectively. The proper angle of flow can be taken as $^{180}\!/_2 = 90°$, and reference to Fig. 182 shows that the factor for second harmonic with $\theta = 90°$ and $\alpha = ^3\!/_2$ is 0.236, and for the direct-current term is 0.146. This makes the direct-current component of the total space current 59.5 ma and the second-harmonic component 96.0 ma. Assuming the d-c grid current to be 15 per cent of the total space current, or 9.0 ma, gives $59.5 - 9 = 50.5$ ma as the direct-current component of the plate current and $96.0 - (2 \times 9.0) = 78$ ma as the second-harmonic component of the plate current. The power supplied by the plate battery is $1000 \times 0.0505 = 50.5$ watts, and the power delivered to the load is $(1000 - 150)0.078/2 = 33$ watts. Hence the plate efficiency is 66 per cent and the plate loss 17.5 watts. The required load impedance between plate and cathode is $(1000 - 150)/0.078 = 10{,}900$ ohms. Substitution in Eq. (174) shows the proper grid bias is 517 volts and the grid exciting voltage is $517 + 120 = 637$ volts. The grid driving power is approximately $637 \times 0.009 = 5.7$ watts.

The accuracy of the above analysis is indicated by the following comparison with results obtained by an exact point-by-point analysis.

	Approximate analysis using Fig. 182	Exact point-by-point analysis
D-c plate current	50.5 ma	51.1 ma
D-c grid current	9.0 ma	8.5 ma
A-c plate current	78 ma	81.6 ma
Power input	50.5 watts	51.1 watts
Power output	33 watts	34.7 watts
Plate loss	17.5 watts	16.4 watts
Plate efficiency	66 per cent	68 per cent

Harmonic Generation by Grid-circuit Distortion.—Another possible method of harmonic generation makes use of the non-linear relationship between the grid current and grid voltage to distort the wave shape of the voltage applied to the grid of the tube and thereby to produce harmonics. The fundamental circuit arrangement of a harmonic generator of this type is illustrated in Fig. 183, in which Z_g is an impedance placed in series with the grid circuit and designed to offer a high impedance to the desired harmonic components of the grid current which flows when the grid is driven positive at the peak of each cycle. Such harmonic generators are capable of developing comparatively large outputs on high-order harmonics, or can be designed to generate a small output on a large number of harmonics simultaneously.[1]

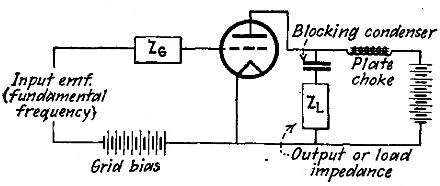


FIG. 183.—Basic circuit of grid-distortion harmonic generator.

[1] For further discussion see F. E. Terman, D. E. Chambers, and E. H. Fisher, Harmonic Generation by Means of Grid Circuit Distortion, *Trans. A.I.E.E.*, vol. 50, p. 811, June, 1931.

64. Linear Amplifiers.—The linear amplifier is a Class C amplifier modified by adjusting to make the output voltage proportional to the exciting voltage. Such amplifiers are used extensively in the amplification of modulated waves since they preserve the modulation without distortion.

The linearity of a linear amplifier can be determined by plotting a curve of output voltage as a function of exciting voltage, as in Fig. 184. If the characteristic curves of the tube were straight lines up to the point where the maximum grid potential exceeded the minimum plate potential, the linear amplifier would be operated with a bias equal to the cut-off value for the plate-supply potential. The plate-current pulses would then be half-sine waves having an amplitude proportional to the exciting voltage, and the amplifier would be perfectly linear up to the flattening off that would occur when the maximum grid potential became greater than the minimum plate potential. The actual characteristic curves of a tube are curved, with the result that, if the bias is the cut-off value, the relationship between output and exciting voltages tends to be nonlinear, as shown by b of Fig. 184. This curvature of the characteristic can be virtually eliminated by biasing the linear amplifier to slightly less than cut-off, giving the result shown by a in Fig. 184. This proper bias is approximately that obtained by extending the straight-line part of the tube characteristic, as shown in Fig. 168a, to give a "projected cut-off" bias.

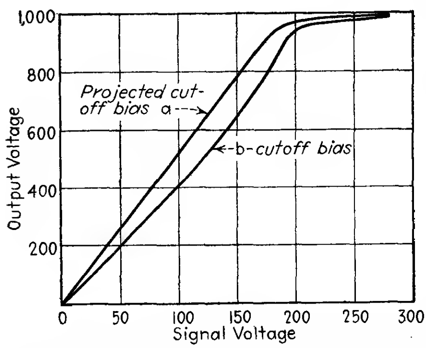


FIG. 184.—Curves showing linearity of linear amplifier for different bias conditions.

The linearity of a linear amplifier also depends upon the load impedance, being greater with load impedances that are high compared with the average effective plate resistance. At the same time, a high load impedance requires a smaller plate-current pulse to make the minimum plate potential become less than the maximum grid voltage, so that the output power over the region of linearity is reduced. The proper load impedance is roughly that which makes the plate loss equal to or slightly greater than the rated dissipation of the tube when the excitation is sufficient to make the minimum plate potential approximately equal to the maximum grid potential.[1] The proper load impedance is hence

[1] It is permissible to make the adjustment so that the plate loss with full exciting voltage is slightly greater than the rated value because, when the signal is a modulated wave, the exciting voltage is at this crest value only a small fraction of the time.

determined by assuming a reasonable[1] peak space current I_m, a suitable combination of maximum grid potential, and minimum plate potential, and then calculating the performance either by the exact or by the approximate method, using the angle of current flow calculated by the following rearrangement of Eq. (167)

$$\cos\left(\frac{\theta}{2}\right) = \frac{1}{1 + \dfrac{\mu E_{\max} + E_{\min}}{\mu E_c - E_b}} \tag{175}$$

In the usual case when the bias is at projected cut-off, the value of θ obtained from Eq. (175) will exceed 180° (*i.e.*, $\cos\left(\frac{\theta}{2}\right)$ will be negative). In case the right-hand side of Eq. (175) is not within the range of +1 to −1, it means that the angle of flow is either 360° or 0.

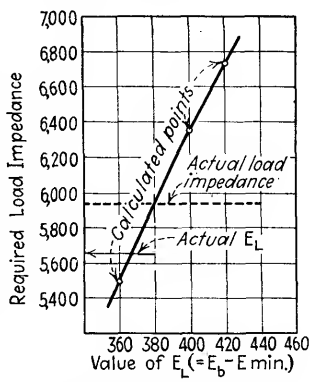


FIG. 185.—Determination of the actual output voltage for a known load impedance by plotting a curve of load impedance required for several assumed values of output voltage.

The calculation of the output voltage of a linear amplifier for a known exciting voltage and given load impedance must be carried out by a cut-and-try process. First an estimate is made of the alternating voltage E_L which might reasonably be expected across the load between plate and cathode. The minimum plate potential is then $E_{\min} = E_b - E_L$. It is now possible to make an exact point-by-point analysis starting with the assumed $E_{\min}$ and the known grid bias and exciting voltages, in this way determining the load impedance that would be required to give the assumed value of E_L. This will generally differ somewhat from the actual load impedance, so that a new estimate must be made of E_L, and the calculation must be repeated to obtain a new load impedance. By repeating this process two or three times it is possible to draw a curve such as shown in Fig. 185 and to interpolate to obtain the output voltage E_L for the load impedance actually employed.

When the relationship between the output voltage and exciting voltage has been obtained either by calculation or experimentally, the distortion of the modulation envelope for any desired degree of modula-

[1] The allowable peak space current in the case of thoriated and oxide-coated cathodes is the same in linear amplifiers as in Class C amplifiers, but with tungsten cathodes the permissible value for linear amplifiers is about two-thirds the peak emission employed with Class C amplification if distortion from incipient saturation is to be avoided.

tion can be obtained with the aid of Eqs. (151) or (152) by knowing the output at certain critical points on the modulation cycle, as illustrated in Fig. 186. If the distortion is to be kept low, it is apparent that the crest exciting voltage at the peak of the modulation cycle must not extend appreciably beyond the point where the linearity curves begin to flatten as a result of an excessively low minimum plate potential.

The plate efficiency of a linear amplifier assuming a straight-line tube characteristic is the same as given by Eq. (166) for Class B audio amplifiers, while, in the case of an actual tube with the bias at projected cut-off, the efficiency is slightly less. With the maximum allowable excitation the plate efficiencies of linear amplifiers normally lie between 50 and 65 per cent, while with signals smaller than the maximum permissible value the plate efficiency will be directly proportional to the signal amplitude. Since the carrier component of a completely modulated wave is exactly half of the crest of the modulation cycle, the efficiency of practical linear amplifiers when the carrier is not modulated is around 30 per cent, and the carrier power that can be developed by a given tube is slightly less than one-fourth of the output power that the same tube could develop when operating as a Class C amplifier. During the periods when the carrier is modulated, the average efficiency increases because the average input over the modulation cycle is the same as without modulation, but the output power has been increased as a result of the side-band energy.[1] The result is that the plate loss in a linear amplifier will be greatest when the carrier is unmodulated, and, in fact, will be about the same during these unmodulated intervals as it would be with an unmodulated exciting voltage having an amplitude equal to the amplitude at the peak of the modulation cycle.

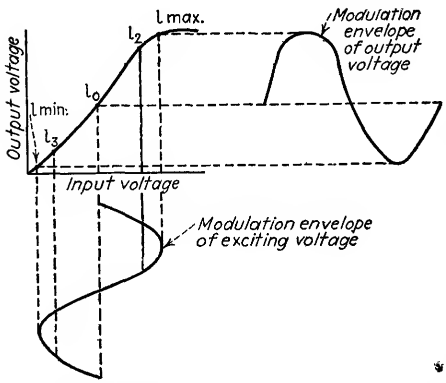


FIG. 186.—Diagram showing how linearity curve can be used to determine distortion of modulation envelope with the aid of Eq. (151) or (152).

Adjustment of Linear Amplifiers.—The procedure to be followed in adjusting linear amplifiers depends upon the circumstances and the measuring equipment available. When ample exciting voltage can be

[1] The average input is substantially independent of modulation because of the fact that the d-c current drawn by the tube is independent of the modulation. The modulation merely superimposes upon the d-c current alternating components that do not represent power with respect to the source of plate voltage.

had, the simplest procedure is to determine experimentally the relationship between the output voltage and exciting voltage for different load impedances until the best compromise is obtained between linearity and power output. When the exciting voltage available is barely sufficient to supply the desired carrier, it is then necessary to modulate this carrier and to make use of a cathode-ray oscillograph or modulation meter to determine the amount of distortion introduced in the modulation envelope by the linear amplifier. The load impedance and the amplitude of the modulated carrier are then varied until the best compromise is obtained between power output and freedom from distortion.

In making adjustments on a linear amplifier, it is to be noted that the grid current which is drawn varies greatly during the modulation cycle. This places a variable-load impedance upon the exciter which tends to distort the exciting voltage. This must be taken into account either by balancing the effects produced by the variable load impedance against other causes of non-linearity or by making the power capacity of the exciter so great that the variation in load will have negligible effect. In some cases a resistance is actually connected across the tank circuit to waste a large amount of power so that the variable power consumed by the following tube will not have much effect.

The bias for linear amplifiers must be constant irrespective of the degree of modulation or the grid current, and it is consequently normally obtained from a fixed source or by self-bias. Grid-leak arrangements are not permissible under any circumstances.

Problems

1. By means of curves of the type shown in Fig. 152 or Fig. 154 demonstrate qualitatively that, when the operating grid bias of a Class A power amplifier is made more negative, the proper load resistance is increased.

2. A certain power tube having $\mu = 3$ and $R_p = 2400$ ohms is to be operated as a Class A power amplifier with $E_B = 350$ volts. Assuming that a reasonable value of E_0 is $0\ 10E_B$, determine the proper load resistance, proper grid bias, and maximum undistorted power, and estimate the plate efficiency that can be expected by assuming $I_{\min}/I_0 = 0.10$.

3. When the plate potential of the tube of Prob. 2 is raised to 500 volts, it is necessary to make the bias -120 volts to prevent excessive plate dissipation. At this operating point the plate resistance is 2800 ohms. Calculate power output and approximate plate efficiency. Compare the results with those of Prob. 2 and explain the reason for the differences.

4. Using a set of $I_p - E_p$ characteristic curves of an actual power tube, determine the following for an operating condition recommended as suitable for Class A power amplification: (*a*) maximum power output; (*b*) second-harmonic distortion; (*c*) actual plate efficiency, taking into account the "rectified" plate current.

NOTE: The curves are preferably a full-page blueprinted reproduction of some actual tube. It is possible, however, to use curves given in tube manuals or the curves of Fig. 75 if nothing better is available.

5. Assuming that $E_0/E_b = 0.1$ and that $I_{\min}/I_0 = 0.1$, plot a curve showing plate efficiency as a function of $\dfrac{\text{load resistance}}{\text{plate resistance}}$ for a Class A amplifier in which the grid is not driven positive.

6. Using a set of $I_p - E_g$ characteristic curves of an actual power-amplifier tube, plot a dynamic characteristic of the type shown in Fig. 152 for an operating condition recommended by the tube manufacturer as suitable for Class A operation.

7. Assuming that E_B and E_0 are fixed and that the plate resistance is constant, plot for a Class A power amplifier, as a function of $\dfrac{\text{load resistance}}{\text{plate resistance}}$, curves showing: (*a*) relative undistorted power output obtainable; (*b*) relative signal voltage required; (*c*) relative power output for 1 volt of signal. Assume the grid is not driven positive.

8. Using a 2A3 tube as a Class A amplifier with the conditions recommended for $E_B = 250$ volts (see Table VIII), calculate the following when the maximum permissible signal (45 volts crest) is applied: (*a*) a-c plate current; (*b*) a-c voltage developed across the load; (*c*) maximum and minimum instantaneous plate voltage and currents reached during the cycle (assuming zero distortion); (*d*) power dissipated in load; (*e*) plate efficiency.

9. It is to be noted from Eq. (153*b*) that driving the grid of a Class A amplifier more positive causes the proper load resistance for fixed bias to become greater. Demonstrate qualitatively that this must be the case from dynamic characteristics of the type shown in Figs. 152 and 156.

10. The grid of a single Class A power tube is excited from a tube having a plate resistance of 10,000 ohms by a transformer in which the ratio of primary to secondary turns is 2. If the crest exciting voltage desired is 60 volts and the instantaneous grid current at the positive peak is 6 ma, calculate the second-, third-, and fourth-harmonic distortion produced in the exciting voltage by the grid current.

11. A particular output transformer is to couple the plate circuit of a tube having $R_p = 2000$ ohms to a load of 600 ohms. The transformer must offer a load impedance of 4000 ohms to the tube and give a response that does not drop to less than 70.7 per cent of the mid-range value between 80 and 6000 cycles.

a. Specify the required turn ratio, the largest permissible leakage inductance, and the lowest permissible primary inductance.

b. If the efficiency is to be at least 80 per cent, specify the maximum allowable primary and secondary resistance, assuming that the loss is divided equally between primary and secondary.

12. An output transformer with the following constants is to be used with a push-pull amplifier employing Type 6F6 tubes operated as Class A triode amplifiers (not Class AB):

Primary inductance	18	henries
Leakage inductance (referred to primary)	0.20	henries
Step-down ratio	30	
Resistance, primary	400	ohms
Resistance, secondary	1.2	ohms

Determine the proper load resistance that should be used, and calculate and plot the frequency response that can be expected when this load resistance is used.

13. Using a set of $E_p - I_p$ characteristics of an actual pentode (or beam) tube, determine the following for a Class A operating condition recommended by the manufacturer: (*a*) maximum power output; (*b*) second- and third-harmonic distortion; (*c*) actual plate efficiency, taking into account the "rectified" plate current.

NOTE: Curves are preferably a full-page blueprinted reproduction for some actual tube, but in the absence of such curves it is possible to use Fig. 163 or characteristics given in a tube manual.

14. From load lines plotted on the characteristic curves of Prob. 13, plot shape of output wave and evaluate output power for the load resistance of Prob. 13 and for load resistances twice as great and half as great.

15. Explain why the individual tubes in a push-pull amplifier must have substantially identical characteristics if the full advantages of the push-pull connection are to be realized.

16. From the composite curves of Fig. 165 determine the power output and the second- and third-harmonic distortion for the three load resistances shown, when the peak signal voltage equals the grid bias.

17. Check the operating conditions specified for Class B operation of the Type 800 tube in Table IX with the aid of Eqs. (164) to (166) and Fig. 75. Do this by starting with the specified load resistance and determining the crest alternating-current voltage across the load. Then check power output and plate efficiency.

18. *a.* Using the characteristic curves of Fig. 169, derive a set of composite curves for Class B operation at the point: plate voltage = 300 volts; control-grid potential = 0. Draw load lines for plate-to-plate resistances of 3000, 5000, and 8000 ohms, and calculate power output and second and third harmonics when the signal potential between grid and cathode is 50 volts peak.

b. Derive and sketch the shape of the plate current wave in the individual tubes, and also the shape of the output wave.

19. Design an audio-frequency amplifier which will deliver an undistorted output of 15 watts when operated from a microphone that has an internal impedance of 600 ohms and which develops a voltage of approximately 100 μv on open circuit. The over-all frequency response should not vary more than 3 db over the frequency range of 80 to 8000 cycles. The load impedance to which the output is delivered is 10 ohms. The design includes selection of tubes, circuit layout, circuit constants, provision for volume control, specification of transformer constants, design of filter systems to prevent regeneration, etc., as well as calculation of expected frequency response. In order to obtain the required frequency response, it is permissible to use equalization provided the constants and performance of the equalizer are specified.

20. Redesign the amplifier of Prob. 19 to provide sufficient negative feedback in the power stage to reduce the distortion otherwise appearing in the output by a factor of 5.

21. Design a Class C amplifier (including tank circuit) employing a Type 852 (or other) tube (see Table IX) when the plate voltage is 2500 and the frequency of operation is 7500 kc. Use tube characteristics from a tube manual and calculate the expected performance, using the approximate method. Include provisions for neutralizing in the design.

22. In the amplifier tube used in the example of Sec. 61, calculate the performance when the following changes are made in the design: (a) plate voltage increased to 1500 volts; (b) angle of flow changed to 90°. In each case make whatever changes in unspecified quantities are necessary to obtain optimum performance while keeping within the allowable plate dissipation and filament emission. Tabulate results of this problem along with those of the example and discuss the differences observed.

23. *a.* If the coupling to the load in a properly adjusted Class C amplifier is removed, the d-c plate current will decrease to a small fraction of its original value, while the d-c grid current will increase somewhat. Explain the reasons for this.

b. If in (*a*) the removal of the load causes only a moderate decrease in d-c plate current (such as 50 per cent), what does this mean?

24. Design a screen-grid Class C amplifier using a Type 865 screen-grid tube (see Fig. 76 and Table IX) operating at a plate potential of 625 volts and a frequency of 4000 kc.

25. Design a harmonic generator for the fourth harmonic using the same tube as in the example of Sec. 63, for angles of flow of (*a*) 45°, (*b*) 67½°, (*c*) 90°, using the approximate method of analysis. Tabulate the results for the three cases together with the results for the second harmonic in the example, and discuss the differences.

26. Design a harmonic generator for the second harmonic, using a Type 852 or other assigned tube. Make calculations by the approximate method, and make use of data in tube manuals to get the necessary tube data.

27. Design a linear amplifier to handle a completely modulated wave using a Type 800 tube of Fig. 75. In this design use the approximate method to determine the proper load impedance and other operating conditions. From the results calculate the approximate carrier output obtainable, and the plate efficiency and plate losses when the carrier is (*a*) unmodulated; (*b*) completely modulated.

28. Derive Eq. (174).

29. Calculate the linearity curve for the amplifier of Prob. 27 and from this determine the second- and third-harmonic distortion produced in the envelope of a completely modulated wave. Neglect the distortion of the exciting voltage caused by the variable exciting power required by the linear amplifier.

CHAPTER VIII

VACUUM-TUBE OSCILLATORS

65. Vacuum-tube Oscillator Circuits.—A vacuum tube is able to act as an oscillator because of its ability to amplify. Since the power required by the input of an amplifier tube is much less than the amplified output, it is possible to make the amplifier supply its input. When

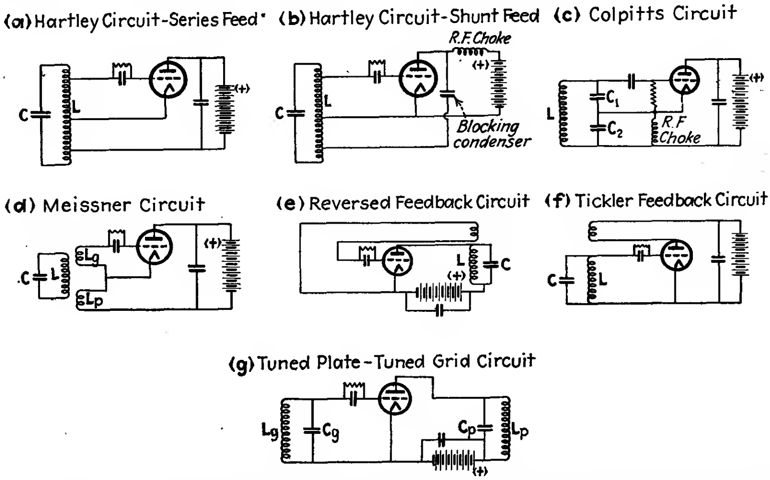


FIG. 187.—Typical oscillating circuits. In each case the frequency is determined by a resonant circuit, and the arrangement is such that the tube acts as an amplifier supplying its own input voltage.

this is done, oscillations will be generated and the tube acts as a power converter that changes the direct-current power supplied to the plate circuit into alternating-current energy in the amplifier output.

Any amplifier circuit that is arranged to supply its own input voltage in the proper magnitude and phase will generate oscillations. Many circuits can be used for this purpose, of which a number are shown in Fig. 187. In general, the voltage fed back from the output and applied to the grid of the tube must be approximately 180° out of phase with the voltage existing across the load impedance in the plate circuit of the amplifier, and must have a magnitude sufficient to produce the output power necessary to develop the input voltage. In the Hartley and Colpitts circuits this is accomplished by applying to the grid a portion of the voltage developed in the resonant circuit. In the Meissner and

the two feedback circuits, mutual inductances are employed, while the tuned plate-tuned grid circuit transfers energy to the grid tuned circuit through the grid-plate tube capacity. All the circuits shown in Fig. 187 except that at Fig. 187*b* use series feed in the plate circuit, but it is possible to employ shunt feed; in fact, the shunt feed is usually preferred in practical cases.

The frequency at which the oscillations occur is the frequency at which the voltage fed back from plate circuit to the grid is of exactly the proper phase and magnitude to enable the tube to supply its own input. In oscillators associated in some way with a resonant circuit, as are all those of Fig. 187, the frequency of oscillation approximates very closely the resonant frequency of this circuit.

66. Design and Adjustment of Power Oscillators.[1]—Oscillators in which the object is to produce appreciable power output are adjusted so that the tube operates as a Class C amplifier. The only difference insofar as voltage and current relations, calculation of performance, etc., are concerned is that, since the oscillator must supply its own excitation, its output is less than that of the corresponding Class C amplifier by the grid driving power.

The real difference between power oscillators and Class C amplifiers is in the circuits, since the oscillator must be arranged to supply its own excitation and must also operate with a grid-leak bias. Furthermore, where stability of frequency is important, additional considerations are involved in the circuit design, as discussed in Sec. 67. In particular the effective Q of the tank circuit must be as high as possible if frequency stability is important, even though this involves some sacrifice in tank-circuit efficiency.

Grid-leak bias is necessary in order to make an oscillator self-starting and to insure stable operation under the desired voltage and current relations. The use of a grid leak makes the oscillator self-starting because, when the electrode voltages are first applied, the grid bias is zero, making the plate current and hence the amplification large. Any thermal agitation or transient voltage of the frequency of the resonant circuit will hence be amplified and thereby start the building up of oscillations. These oscillations cause the grid to draw current which biases the grid negative as a result of the grid-leak resistance. This reduces the d-c plate current and hence the amplification of the tube, until ultimately an equilibrium is established at an amplitude such that the plate current is reduced to the point where the amplification is exactly one, *i.e.*, where the power generated in the output is just able to sustain

[1] An excellent and exhaustive treatment of vacuum-tube oscillators is to be found in the series of articles by D. C. Prince, Vacuum Tubes as Power Oscillators, *Proc. I.R.E.*, vol. 11, pp. 275, 405, 527, June, August, October, 1923.

the amplitude of oscillations required to produce this power, and where there is no surplus remaining to cause further increase in amplitude. This equilibrium normally occurs when the minimum plate potential is quite small, approximating the values commonly considered desirable in Class C amplifier operation. When a fixed grid bias is used with Class C operation, oscillations cannot build up when the plate voltage is applied because the grid bias is greater than cut-off.

In order to achieve stability, it is necessary that a decrease in the amplitude of the oscillations increase the amplification, for otherwise a minute irregularity producing a small decrease in amplitude will cause the amplification to become less than the equilibrium value of unity. This results in further progressive decreases in amplitude and the ultimate dying out of the oscillations. The grid leak ordinarily meets the requirement for stability because any decrease in the amplitude of oscillations also reduces the bias developed by the grid-leak arrangement, thereby tending to maintain the same maximum grid potential even with less excitation. In contrast with this, reduction in amplitude with fixed bias reduces the amount the grid is driven positive, which tends to reduce the output and produce instability.

The design of a power oscillator, as far as the operating conditions of the tube are concerned, can be carried out on paper exactly as for a Class C amplifier. Among other things, this determines the alternating grid and plate voltages desired. The ratio of these fixes the relative coupling from the tank circuit to the grid and plate circuits, while the absolute magnitude of the couplings must be such that, with the expected power output delivered to a tank circuit having the appropriate effective Q, the voltages will have the proper magnitude. The details involved are made clear by the example given below. The grid condenser must be large enough to have a low reactance compared with the grid-leak resistance, and it should be at least five to ten times the grid-cathode tube capacity. At the same time the capacity must not be so large as to cause intermittent oscillations as discussed below. When shunt feed is employed, the inductance of the shunt-feed choke is effectively in parallel with a portion of the tank circuit; in order to keep the circulating current and hence losses in the choke low, the choke inductance should be at least five to ten times the inductance of the tank circuit.

Example.—An oscillator is to be designed to operate under the same conditions as the Class C amplifier of Sec. 61, using a Hartley circuit having an effective Q of 50. Assuming the grid and plate taps are to be at the ends of the tank-circuit inductance, the total alternating voltage across the tank circuit will be the sum of the exciting and alternating plate voltages, or 254 + 850 = 1104 volts. The filament tap is located so that the ratio between grid and plate voltages is $^{254}/_{850}$. The tank-circuit inductive reactance ωL required is found by Eq. (172) to be $(1104/\sqrt{2})^2/(56 \times 50) = 218$ ohms for a tank-circuit power of 56 watts. From this the required tank-circuit induc-

tance and capacity for any frequency can be calculated. The tank-circuit circulating current is $^{1104}/_{218} = 5.06$ amp. crest. Since the grid bias required is 134 volts, with an estimated grid current of 0.0135 amp., the grid-leak resistance can be estimated as $134/0.0135 = 10{,}000$ ohms. The grid condenser capacity should be large enough to be an effective short circuit to the grid-leak resistance at the operating frequency, but not so large as to cause intermittent oscillations (see below).

If the grid and plate connections were not made to the ends of the coil, the voltage across the tank circuit would have to be increased accordingly and a higher tank-circuit inductance would be required to obtain the required conditions with the same effective circuit Q. If other oscillator circuits than the Hartley had been employed, minor modifications to this general procedure would be necessary.

Adjustment of Power Oscillators.—The method to be followed in adjusting a power oscillator depends upon the completeness with which the paper design has been carried out. When the design has been worked out to the point where the couplings of grid and plate electrodes to the tank circuit have been determined, as well as the coupling required between the load impedance and the tank circuit to give the desired Q, and when these appropriate couplings have all been realized by means of measurements or calculations, the adjustment is very simple. One starts with a grid-leak resistance that will give the desired bias with the estimated grid current, notes the grid bias actually obtained as the oscillator operates, and readjusts the grid-leak resistance until the desired bias is realized. The tube will then be operating under the design conditions.

More commonly the design is carried only to the point where the desired operating conditions for the tube have been determined, as well as the required tank-circuit inductance, capacity, and current. It is then necessary to adjust the load coupling; the grid, plate, and filament taps (or couplings); and the grid-leak resistance by trial until the desired operating conditions are realized. The simplest procedure for doing this is to place a thermocouple ammeter in series with the tank circuit, make the coupling between the load and tank circuit small, and use a grid-leak resistance that will give the desired bias on the basis of the estimated grid current. The coupling between the grid and tank circuits is then set by guesswork to a reasonable value, and the coupling from tank to plate circuits is varied until the expected tank-circuit current is obtained. Next, the coupling to the load is increased until the d-c plate current is as close as possible to the expected value, after which the input power, output power, plate efficiency, and grid bias are noted. The coupling from tank circuit to grid circuit is then adjusted to realize more nearly the desired conditions, keeping in mind that closer coupling to the grid will increase the maximum grid potential but will reduce the angle of flow, as explained below. After the best adjustment of the grid excitation has been obtained, the plate coupling, load coupling, and

grid coupling are readjusted in sequence over and over again, always working toward the desired operating conditions. As these begin to be approached, it is also necessary to change the grid-leak resistance at the end of each sequence in order to approach the desired grid bias.

If an ammeter is not used to read the tank-circuit current, it is then necessary to start by arbitrarily selecting a reasonable coupling between the tank and plate circuits, after which the proper excitation, load coupling, and grid-leak resistance are derived as above. If the final results obtained in this way make the tank-circuit loss too great or the effective Q of the tank circuit too low, then it is necessary to change the coupling between tank and plate circuits and repeat the entire adjustments.

The trial-and-error adjustment of oscillators can be systematized by keeping in mind the fact that oscillators adjusted for reasonable efficiency have the following characteristics: (1) The bias is determined primarily by the exciting voltage because the grid current normally increases so rapidly as the maximum grid potential is increased that the bias automatically adjusts itself so that the grid is driven moderately but not excessively positive. Increasing the excitation accordingly increases the bias and reduces the angle of flow while at the same time increasing the maximum grid potential somewhat. (2) The maximum positive grid potential E_{max} is determined both by the exciting voltage and by the grid-leak resistance, being greater as the grid-leak resistance is reduced or as the excitation is increased. (3) The tank-circuit current is determined primarily by the coupling between the tank circuit and plate circuit of the tube, since the tank current is normally such that the crest alternating voltage developed between plate and cathode is just less than the plate-supply voltage. Hence more coupling between the plate and tank circuits reduces the tank-circuit current. (4) The minimum plate potential E_{min} is determined largely by the coupling between the load and the tank circuit, as is also the d-c plate current. Close coupling increases the dissipation in the load for a given tank-circuit current. This tends to reduce the amplitude of oscillations, thereby increasing the minimum plate potential and increasing the d-c plate current which the plate draws. (5) The grid current is determined by the exciting voltage, grid-leak resistance, and load coupling, and for practical purposes is fixed by the maximum grid voltage E_{max} and the minimum plate potential E_{min}.

Factors Controlling Amplitude of Oscillation.—The factors affecting the amplitude of oscillation of a properly adjusted power oscillator are the load coupling, the anode voltage, and the frequency. With proper adjustment, equilibrium is always obtained with an amplitude such that the minimum plate potential is small compared with the plate-supply

voltage. This condition is maintained under widely varying conditions because the d-c plate current is quite sensitive to changes in the minimum plate potential E_{min}, and hence causes large changes in power output for small changes in amplitude of oscillations. Thus, if the plate-supply potential is varied, the alternating plate-cathode voltage is changed almost in proportion. Likewise, if the frequency is changed by varying the capacity of the tank circuit, the current in the tank circuit will be almost exactly inversely proportional to frequency because the voltage across the tank circuit tends to be constant even if the inductive reactance ωL of the tank circuit varies with frequency.

Varying the resistance of the tuned circuit has relatively little effect on the amplitude of the oscillations but does change the d-c plate current. This is a result of the fact that, when the resistance of the resonant circuit is increased, it is impossible for the oscillations to maintain their amplitude, inasmuch as the original oscillating current in flowing through the added resistance causes more energy to be consumed in the resonant circuit than is supplied from the plate-voltage source. The result is that immediately upon the insertion of additional resistance the amplitude of the oscillations becomes less. This makes the minimum plate voltage larger, increasing the amplitude of the plate-current impulses and resulting in the resonant circuit receiving additional energy. The amplitude of oscillations then assumes a new equilibrium point at which the enlarged plate-current impulses supply sufficient energy to the resonant circuit to maintain the oscillations with the higher resistance circuit. Inasmuch as the amplitude of the alternating plate-cathode voltage need decrease only a small percentage in order to increase greatly the amplitude of the plate-current impulses, the main result of adding resistance to the oscillating circuit is to increase the direct current drawn from the plate-supply source, while the amplitude of the oscillations, the grid exciting voltage, and the grid bias are affected relatively much less.

Intermittent Operation.—When a grid leak is used to develop the grid bias, it is sometimes found that the oscillations are periodically interrupted. These interruptions may be at an audible rate or they may be at a radio frequency. They are the result of an excessively large time constant R_gC_g for the grid-leak grid-condenser combination, and they can be cured either by decreasing the grid-leak resistance or by reducing the grid-condenser capacity, or both.

Intermittent operation arises from the fact that, when the time constant R_gC_g of the leak-condenser combination is large, the bias voltage across the leak adjusts itself slowly to sudden changes in the amplitude of oscillation. If this rate of adjustment is so slow that oscillations can die out before the bias voltage can change appreciably, then with sudden changes in amplitude the action is very much as though one had a fixed

bias; and as already explained, this leads to instability. The process involved in intermittent operation is therefore as follows: The oscillations first build up in amplitude to the equilibrium condition. Any slight irregularity tending to reduce the amplitude of oscillations will then cause the oscillations to die out because, with the large time constant R_gC_g, the bias tends to remain constant, whereas, in order to prevent the oscillations from dying out, it is necessary for the bias to reduce as the amplitude of oscillations reduces. After cessation of oscillations the grid condenser gradually discharges through the grid leak, reducing the bias until the tube will again amplify. An oscillation thereupon builds up again to the equilibrium value, to repeat the process.

Blocking.—The phenomenon of blocking appears as a sudden stoppage of oscillations, accompanied by a reversal of grid current and an increase of plate current to a value much higher than can be obtained with the full direct-current supply voltage and zero grid potential. A high-power tube is usually destroyed by blocking, since the energy dissipated at the plate is enormous. Blocking is caused by operating conditions that permit secondary electron emission to take place at the grid to such an extent that the grid loses more electrons by secondary emission than it gains from the cathode by direct flow. This causes a reversal of the grid current, resulting in the development of a positive grid-bias voltage by the grid leak, and consequently an excessive plate current flows. In order that blocking may exist, it is necessary that the minimum instantaneous plate voltage and the maximum instantaneous grid potential obtained during the cycle both be high and that the grid leak have a high resistance. Blocking, when it occurs, is the result of attempting to force the output of the oscillator by increasing the load resistance coupled into the plate circuit. This results in a reduction in the amplitude of oscillations, which increases the minimum plate voltage. If the grid excitation is then increased, the secondary electron emission at the grid will be increased because of the increased positive potential reached by the grid. Under unfavorable conditions this will cause the net grid current to become less, which reduces the grid bias and makes the maximum positive grid potential still greater, causing a further reduction in the grid current, and so on.

In order for blocking to occur, the grid must emit secondary electrons. In modern air-cooled tubes the grid structures have usually been so treated that the grid current will never reverse under conditions encountered in ordinary operation, so that blocking cannot occur in these tubes. However, with water-cooled tubes, or where very high electrode voltages are employed, it is not always possible to construct the tube in such a way that negative grid current can be avoided. Such tubes are therefore susceptible to blocking, and either must be operated with considerable

care or must be provided with a rectifier arrangement as shown in Fig. 188, which has the resistance R adjusted so that the rectified current always exceeds the most negative grid current obtainable. In this way the current through the grid leak can never reverse, and all possibility of blocking is avoided.

Tubes for Power Oscillators.—The same tubes are used for power oscillators as are employed for triode Class C power amplifiers (see Sec. 62). For output powers up to about 1 kw. this means that air-cooled tubes are used, while water-cooled tubes are employed where greater power is required.

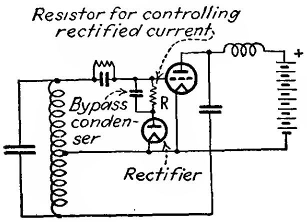


FIG. 188.—Oscillator provided with grid rectifier to prevent the possibility of a negative grid current, with the resulting possibility of blocking.

67. Frequency and Frequency Stability of Generated Oscillations.— The alternating current generated by the vacuum-tube oscillator has a frequency such that the voltage which the oscillations apply to the grid of the tube is of exactly the proper phase to produce the oscillations that supply the required grid exciting voltage. This approximates the resonant frequency of the tuned circuit, but the exact value is also usually influenced by such factors as the effective Q of the tuned circuit, the electrode voltages of the tube, the load coupling, etc.

For many requirements it is essential that the generated frequency be as nearly constant as possible over both short and long time intervals. The first step in achieving this is to maintain the resonant frequency of the tank-circuit constant. Factors that can cause the resonant frequency to change are aging, variation of inductance and capacity with temperature, variations in the interelectrode capacity of the oscillator tube as a result of changes in temperature or tube replacements, and variations in the reactance that the load couples into the tuned circuit. Temperature effects are particularly troublesome because the energy dissipated in the circuit causes progressive heating that is very difficult to compensate for. The temperature coefficient of tuned circuits depends upon the construction, but is commonly in the range 10 to 100 parts in a million per degree centigrade.

Constancy of the resonant frequency of the tank circuit does not insure a stable frequency, however, since the frequency actually generated normally differs from the resonant frequency of the tuned circuit by an amount that depends upon the resistance of the tank circuit, the load resistance coupled into the tank circuit, and the plate-supply, bias, and filament voltages. This can be understood by considering what would result if the oscillations were at the resonant frequency of the tuned circuit in some typical case. Thus take the Colpitts circuit of Fig. 189*a*

and simplify by assuming that the shunt-feed choke has infinite reactance, that the grid condenser has negligible reactance, that the grid current is negligibly small, and that the grid-leak resistance is infinitely large. The only factors affecting the frequency are then the tank-circuit resistance and the coupled load resistance. This leads to the equivalent circuit of Fig. 189*b*, which leads to the vector diagram of Fig. 189*c* when the frequency is the resonant frequency of the tank circuit. Here $-\mu E_{gf}$ represents the effective voltage acting in series with the plate circuit of the tube. Since the tank circuit offers a resistance load at its resonant frequency, the voltage E_{pf} across the tank circuit is in phase with $-\mu E_{gf}$. The current i_L flowing through the inductive branch of the tuned circuit lags E_{pf} by an angle α that is slightly less than 90° as a result of the resistance in the coil and whatever load resistance is coupled into the

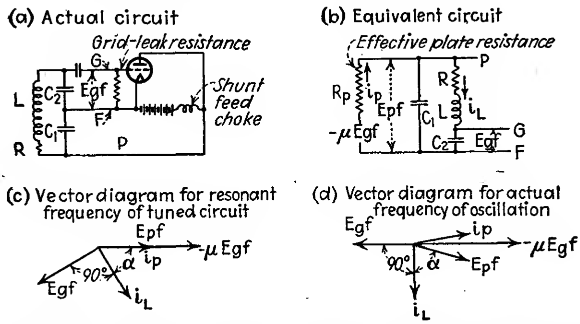


FIG. 189.—Circuit of Colpitts oscillator, together with simplified circuit, and vector diagrams showing how tank circuit resistance makes it necessary for the actual frequency to differ slightly from the resonant frequency.

coil inductance. This current also flows through the condenser C_2 and produces the voltage E_{gf}, which lags the current i_L by 90° and is applied to the grid of the tube. This grid voltage should be such as to produce the voltage $-\mu E_{gf}$ acting in series with the plate, but it fails to do so because the frequency was assumed to be the resonant frequency of the tank circuit.

In order for the voltage on the grid to have exactly the required phase (*i.e.*, for vector E_{gf} and $-\mu E_{gf}$ to be in phase opposition), it is necessary for the frequency of oscillation to be slightly higher than the resonant frequency of the tuned circuit, as is the case in Fig. 189*d*. The amount of this frequency deviation obviously depends upon the resistance of the tuned circuit, and hence will change (with a resulting variation in frequency) as the resistance of the circuit is varied, or as the electrode voltages on the tube alter the effective plate resistance. If, in addition, the effect of the grid current, grid-leak resistance, grid-condenser capacity, shunt-feed choke, etc., are taken into account, the

situation is more complicated, but the results are of the same general character.

It is apparent from the above discussion that the frequency stability will be highest when the oscillations tend naturally to occur at a frequency that differs as little as possible from the frequency of the tuned circuit. Under these conditions the things that affect the difference between the actual and resonant frequency will have proportionately less effect upon the frequency.

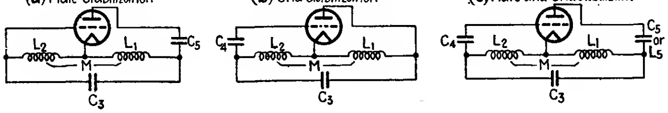
The most important factors contributing to the stability of frequency with variations in electrode voltages, particularly plate supply and filament voltages, are the effective Q of the tank circuit, the constants of the circuits associated with the oscillator tube, and operation so that the power output is relatively small. A high tank-circuit effective Q is important because the frequency change required to compensate for the phase shift resulting from a change in effective plate resistance, etc., is inversely proportional to Q, with the result that the frequency stability is directly proportional to the tank-circuit Q. At the same time, use of a high tank-circuit Q has the disadvantage that it means the load must be loosely coupled, thereby lowering the tank-circuit efficiency and reducing the power output. This, together with the fact that frequency stability is improved by operating the tube so that the alternating currents flowing to the grid and plate electrodes are small (*i.e.*, operation with low power output) means that high frequency stability and large power output are mutually incompatible. In the Hartley, Meissener, and other circuits in which inductive coupling is employed, it is also helpful to use the largest possible coefficients of coupling between the grid and plate coils of the tank circuit in order to reduce the leakage inductances that would otherwise produce phase shifts between the grid and plate alternating voltages.

The frequency of oscillation can be made to approach the resonant frequency of the tuned circuit by inserting suitable reactances in series with the grid or plate electrodes or both.[1] Analysis shows that the impedance required is to a first approximation independent of the electrode voltages of the tube, so that by the use of such compensating reactances it is possible to make the frequency largely independent of the tube conditions. Figure 190 shows how these impedances are inserted under various conditions, and also gives the equations that determine the magnitude of the reactances for the ideal case where the effective tank circuit Q is extremely high. In this ideal case the compensation is perfect and the frequency is absolutely independent of tube voltages. In an actual case the compensating reactances must

[1] For an excellent discussion of such oscillators, see F. B. Llewellyn, Constant-frequency Oscillators, *Proc. I.R.E.*, vol. 19, p. 2063, December, 1931.

HARTLEY OSCILLATOR

(a) Plate Stabilization (b) Grid Stabilization (c) Plate and Grid Stabilization



(a) $C_5 = C_3\left[\frac{L_0}{L_1 + L_2A^2 - 2MA}\right]$

$L_0 = L_1 + L_2 + 2M$

$A = \frac{L_1 + M}{L_2 + M}$

(b) $C_4 = C_3A^2\left[\frac{L_0}{L_1 + L_2A^2 - 2MA}\right]$

$L_0 = L_1 + L_2 + 2M$

$A = \frac{L_1 + M}{L_2 + M}$

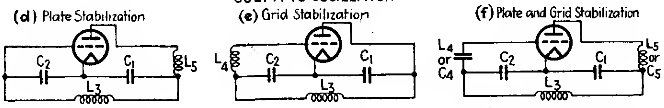
(c) $\frac{1}{C_5} + \frac{A^2}{C_4} = \frac{1}{C_3}\left[\frac{L_1 + L_2A^2 - 2MA}{L_0}\right]$

$L_5 = L_0\frac{C_3}{C_4}A^2 - L_1 - L_2A^2 + 2MA$

$L_0 = L_1 + L_2 + 2M;\ A = \frac{L_1 + M}{L_2 + M}$

COLPITTS OSCILLATOR

(d) Plate Stabilization (e) Grid Stabilization (f) Plate and Grid Stabilization



(d) $L_5 = L_3\frac{C_2}{C_1}$

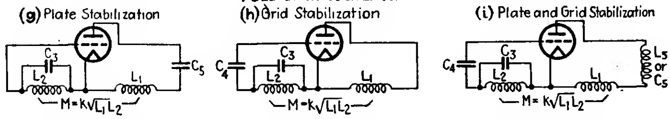
(e) $L_4 = L_3\frac{C_1}{C_2}$

(f) $L_4\left(\frac{C_2}{C_1}\right) + L_5\left(\frac{C_1}{C_2}\right) \doteq L_3$

$L_5 = L_3\frac{C_2}{C_1}\left[1 + \frac{C_2}{C_4}\left(\frac{C_2}{C_1 + C_2}\right)\right]$

FEED-BACK OSCILLATOR

(g) Plate Stabilization (h) Grid Stabilization (i) Plate and Grid Stabilization



(g) $C_5 = C_3\frac{L_2}{L_1}\left(\frac{1}{1 - k^2}\right)$

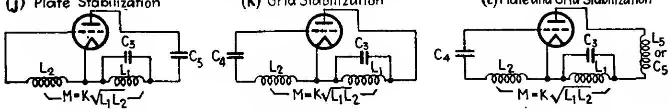
(h) $C_4 = C_3\left(\frac{k^2}{1 - k^2}\right)$

(i) $L_5 = L_1\left[k^2\left(1 + \frac{C_3}{C_4}\right) - 1\right]$

$C_5 = C_3\frac{L_2}{L_1}\left[\frac{1}{1 - k^2\left(1 + \frac{C_3}{C_4}\right)}\right]$

REVERSED FEED-BACK OSCILLATOR

(j) Plate Stabilization (k) Grid Stabilization (l) Plate and Grid Stabilization



(j) $C_5 = C_3\left(\frac{k^2}{1 - k^2}\right)$

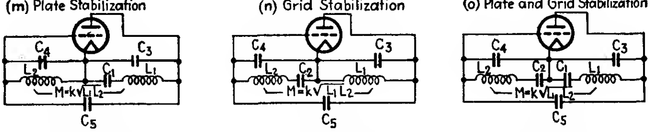
(k) $C_4 = C_3\frac{L_1}{L_2}\left(\frac{1}{1 - k^2}\right)$

(l) $L_5 = L_1\left[1 + \frac{1}{k^2}\left(\frac{L_1C_3}{L_2C_4} - 1\right)\right]$

$C_5 = \frac{C_3}{\frac{1}{k^2}\left(1 - \frac{L_1C_3}{L_2C_4}\right) - 1}$

UNITY COUPLED OSCILLATOR

(m) Plate Stabilization (n) Grid Stabilization (o) Plate and Grid Stabilization

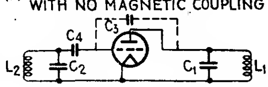


(m) $C_1 = \left(\frac{L_2}{L_1}\right)\left[\frac{C_3k^2\left(\frac{L_1}{L_2}\right) + C_4 + C_5\left(1 + k\sqrt{\frac{L_1}{L_2}}\right)^2}{1 - k^2}\right]$

(n) $C_2 = \left(\frac{L_1}{L_2}\right)\left[\frac{C_4k^2\left(\frac{L_2}{L_1}\right) + C_3 + C_5\left(1 + k\sqrt{\frac{L_2}{L}}\right)^2}{1 - k^2}\right]$

(o) $\omega^2k^2L_1L_2 = \left(\omega L_1 - \frac{1}{\omega C_1}\right)\left(\omega L_2 - \frac{1}{\omega C_2}\right)$

(p) TUNED PLATE-TUNED GRID OSCILLATOR WITH NO MAGNETIC COUPLING

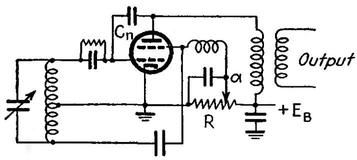


$C_1 = \frac{L_2}{L_1}\left[C_2 + \frac{(1 + \mu)C_3C_4}{C_4 + (1 + \mu)C_3}\right] - C_3$

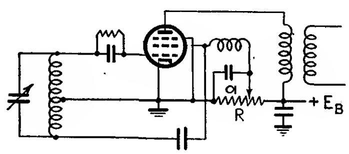
FIG. 190.—Circuits and circuit proportions that will make the generated frequency independent of the tube constants, provided the Q of the resonant circuit is very high (theoretically infinite).

be adjusted by trial, and the compensation, although not perfect, still represents a marked improvement in independence of frequency from tube variations. It will be noted that in some cases the required reactance can be supplied by a suitable choice of grid condenser or plate blocking condenser capacity.

Master-oscillator—Power amplifier and Electron-coupled Arrangements for Realizing High Frequency Stability.—The difficulty of obtaining appreciable power output with high frequency stability has led to the use of master-oscillator power-amplifier arrangements. Here an oscillator designed to have high frequency stability is used to excite the grid of a Class C power amplifier. In this way the oscillator operates with the lowest possible load and hence under conditions that are highly favorable for frequency stability, while the power amplifier is utilized to develop the output power required. It is essential that the Class C amplifier be operated so that the grid current it draws is small. This is because the grid current depends upon the load impedance in the plate circuit of the amplifier, and, if the grid current is not small, variations in the output circuits of the amplifier will produce appreciable reaction upon the oscillator. When a master-oscillator power-amplifier arrangement is properly designed and is operated so that the tank circuit is either at constant temperature or is compensated so that it has the same resonant frequency over ordinary temperature ranges, a frequency stability of the order of 20 to 100 parts per million can readily be maintained.



(a) Neutralized electron-coupled oscillator



(b) Pentode electron-coupled oscillator

FIG. 191.—Electron-coupled oscillator circuits.

The master oscillator and power amplifier can be combined into one tube by using a screen-grid or pentode tube as illustrated in Fig. 191.[1] Here the cathode, control grid, and screen grid are operated as a triode oscillator with the screen serving as the ordinary anode. Only a few electrons are intercepted by the screen, but these are enough to maintain the oscillations. The remaining electrons, which represent most of the space current, go on to the plate and produce the power output by flowing through the load impedance that is connected in series with the plate electrode. This plate current is controlled by the oscillator portion

[1] For further information on such oscillators, see J. B. Dow, A Recent Development in Vacuum-tube Oscillator Circuits, *Proc. I.R.E.*, vol. 19, p. 2095. December, 1931.

of the tube, but, since the plate current of a pentode or screen-grid tube is independent of the plate potential (and hence of the load impedance in the plate circuit) if the minimum plate potential is not too low, there is no reaction between the output circuit and the oscillator section of the tube.[1] This arrangement is called an electron-coupled circuit because the oscillator and the output circuit are coupled by the electron stream.

In the case of screen-grid electron-coupled oscillators, increasing the plate voltage causes the frequency to vary in one way while increasing the screen voltage causes the frequency to vary in the opposite direction. Hence by obtaining the screen potential from a voltage divider, as shown in Fig. 191*a*, and by locating the screen tap at the proper point (as determined by trial), it is possible to make the frequency independent of the plate-supply voltage.

Oscillators Employing Resonant Lines.—In this type of oscillator the resonant circuit that controls the frequency is supplied by a resonant transmission line by taking advantage of the fact that a transmission line short-circuited at the receiving end and an odd number of quarter wave lengths long acts as a parallel resonant circuit. When the frequency is sufficiently high so that the length required is physically short, it is possible by suitable design to obtain circuit Q's ranging from 1000 to 100,000, with values of 10,000 quite practicable under most circumstances. The line can be of the two-wire or of the concentric-conductor type, with the latter having the advantage of no external field and zero radiation losses, but with the disadvantage that adjustments are more difficult. For maximum Q the conductors must be properly proportioned with respect to their spacings. Furthermore the maximum circuit Q is proportional to the square root of the frequency, is independent of the number of quarter wave lengths in the line, and is directly proportional to the size of conductor.[2]

The very high Q obtained by the use of resonant lines is a result of the fact that this construction minimizes skin effect by arranging the surface of the conductor parallel with the magnetic flux lines. The increase

[1] In the case of screen-grid tubes it is necessary to neutralize the electrostatic coupling between plate and the screen, as shown in Fig. 191, if reaction through electrostatic coupling is to be avoided. The same is true of pentodes other than radio-frequency pentodes.

[2] A detailed discussion of the properties of resonant lines when used as tuned circuits is given by F. E. Terman, Resonant Lines in Radio Circuits, *Elec. Eng.*, vol. 53, p. 1046, July, 1934. In this paper it is shown that the circuit Q is given by the relation

$$Q = A\sqrt{f}\,b \tag{176}$$

where

f = frequency in cycles

b = inner radius of the outer conductor in the case of a concentric line, or the

in Q as the frequency increases results from the fact that, although the skin-effect resistance is proportional to the square root of frequency, the length of line required for a quarter wave length is inversely proportional to frequency, so that, as the frequency is increased, the length reduces faster than the skin effect increases. Resonant lines are hence particularly suitable for use at extremely high frequencies, since then the Q is high and the physical dimensions are small.

A practical oscillator circuit utilizing a resonant transmission line to control the frequency is shown in Fig. 193. This is a tuned-plate tuned-grid type of oscillator with the grid tuned circuit supplied by the high-Q resonant line and the plate tuned circuit supplied by a conventional coil and condenser arrangement to which the load is coupled. In an arrangement of this sort the grid tuned circuit controls the frequency, and, because of the extremely high Q obtained with the resonant line, the frequency is substantially independent of ordinary variations in supply voltages or load coupling. For maximum frequency stability the resonant line must be loosely coupled to the grid of the tube in order that the grid losses will be relatively small compared with the circulating energy in the line. This is accomplished by connecting the grid to the line at a point relatively close to the shorted end.[1]

At frequencies above 10 to 20 mc the frequency stability of a resonant-line oscillator is of the same order of magnitude as obtainable with

spacing between conductors in the case of a two-wire line, measured in centimeters

A = a constant depending upon the line proportions and given in Fig. 192.

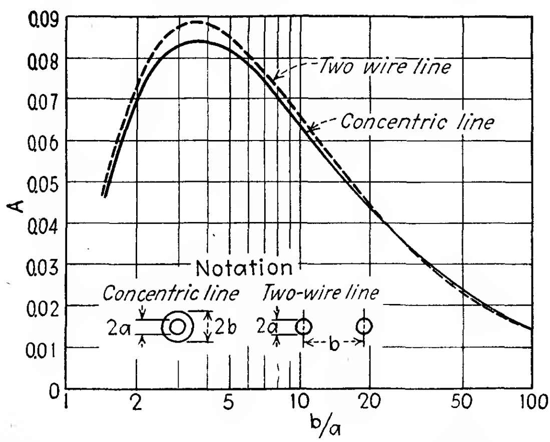


FIG. 192.—Value of A for use in Eq. (176). This neglects radiation from line.

[1] This and other circuit arrangements developed by RCA for utilizing resonant lines to control the frequency of power oscillator are described by C. W. Hansell, Resonant Lines for Frequency Control, *Elec. Eng.*, vol. 54, p. 852, August, 1935.

crystal control if the line is provided with means to maintain the line length constant with variations in room temperature. At the same time the resonant line has the advantage of simplicity, since large amounts of power can be generated by the oscillator tubes and delivered directly to the load without the necessity of power amplifiers.

Oscillators with More Than One Resonant Frequency.—When the electrical network associated with an oscillator tube involves more than one tuned circuit, it may be possible to have two frequencies of oscillation. When these two frequencies are widely different, as is commonly the case when one is dealing with parasitic oscillations, it will ordinarily be found that both oscillations will exist simultaneously, with the high-frequency oscillation being modulated by the low-frequency oscillation. On the other hand, if the two possible frequencies of oscillation are quite close together, it is usually found that only one oscillation exists at a time, since the one that gets established first suppresses the other.

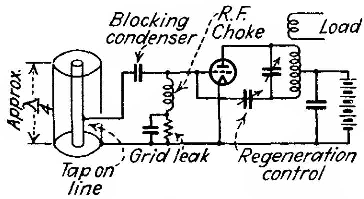


FIG. 193.—Circuit of practical vacuum-tube oscillator employing resonant line to control frequency.

A particularly interesting and important case involving two resonant circuits arises when the load is supplied by a tuned circuit that is coupled to the tank circuit of the oscillator. If the coupling between the two tuned circuits is sufficiently close, there are two possible oscillation frequencies corresponding to the two humps in the coupling curve discussed in Sec. 17. The particular oscillation obtained will then be a matter of chance, and momentary stoppage of oscillations or merely a transient may cause the frequency to jump to the other possible value. The remedy for this situation is either to detune one of the circuits in such a way as to discriminate against one of the possible oscillations or to reduce the coupling to the point where the volt-amperes in the secondary are less than the volt-amperes in the primary. This latter condition is obtained when the coefficient of coupling between the two oscillator circuits is less than the reciprocal of the effective Q of the secondary.[1]

Frequency Range of Oscillators.—The lowest frequency obtainable from an oscillator is limited only by the fact that, as the resonant frequency of the tank circuits is lowered, there is increasing difficulty in obtaining a reasonable circuit Q with reasonable physical dimensions. Frequencies of the order of 5 to 15 cycles a second can, however, be readily obtained with tuned circuits employing iron-cored coils.

[1] Further discussion of problems involved in oscillators employing two circuits tuned to the same frequency is given by D. C. Prince, *loc. cit.*

The high-frequency limit of large tubes is set by the resonant frequency obtained when the capacity of the tank circuit consists only of the interelectrode capacity of the tube and when the inductance is the inductance of the shortest possible leads between electrodes. In the case of tubes physically small, such as the "acorn" tubes, the frequency limit is so high that the low input resistance resulting from the finite transit time is a further limit. With suitable tubes considerable power can be readily obtained at frequencies of 150 to 300 mc, and oscillations at frequencies exceeding 1000 mc can be obtained with "acorn" tubes. The plate efficiency and hence power output always decrease rapidly as the high-frequency limit of a tube is approached, and this, coupled with the fact that tubes for generating the highest frequencies must be physically small, means that very little power is obtainable at the highest frequencies.[1]

68. Parasitic Oscillations.[2]—The term *parasitic* is applied to any undesired oscillation occurring in an oscillator or power amplifier. Such oscillations are commonly encountered when large tubes are employed; they are to be avoided because they absorb power that would otherwise go to generating useful output and because they also often produce excessive voltage stresses in portions of the circuit. In addition, parasitic oscillations will normally cause distortion in linear amplifiers, modulators, and Class A and Class B audio amplifiers. Low-frequency parasitic oscillations will also modulate a carrier wave either at an audio or at a low radio frequency. A rough note from the output also nearly always indicates parasitic oscillations.

Examples of Parasitic Oscillations.—Parasitic oscillations result from the fact that, when the tube capacities, lead inductance, shunt-feed chokes, etc., are all taken into account, it is usually found that several possible modes of oscillation can exist in addition to the desired type of operation. The nature of these parasitic circuits and the means for preventing them from giving rise to oscillations can be understood by considering several typical examples. Consider first the simple Class C amplifier circuit of Fig. 194*a*. At very high frequencies this reduces to the circuit of Fig. 194*b*, which is a tuned-plate tuned-grid oscillator circuit with the grid and plate tuning capacities supplied by the grid-filament and plate-filament tube capacities, respectively. The grid inductance L_1 and plate inductance L_2 are supplied by the inductance of the leads between these two electrodes and ground through the tuning

[1] For further discussion of oscillators for very high frequencies, see M. J. Kelly and A. L. Samuel, Vacuum Tubes as High-frequency Oscillators, *Elec. Eng.*, vol. 53, p. 1504, November, 1934.

[2] For further information see G. W. Fyler, Parasites and Instability in Radio Transmitters, *Proc. I.R.E.*, vol. 23, p. 985, September, 1935.

capacities C_1 and C_2, which can be considered as short circuits at the parasitic frequency. It will be noted that the neutralization is not effective because the coil L_p does not participate in the parasitic oscillation. Instead, the neutralizing condenser in series with the inductance of its leads merely forms a shunt across L_1 to ground through C_2, modifying the resonant frequency of the grid tuned circuit. The circuit shown at Fig. 194*b* will oscillate provided the plate tuned circuit offers inductive

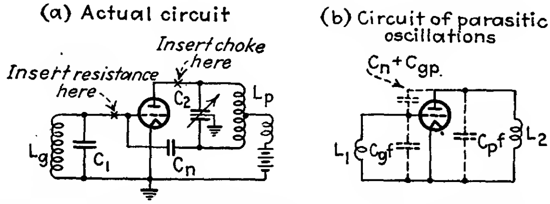


FIG. 194.—Typical Class C amplifier circuit, showing how high-frequency parasitic oscillations are possible.

reactance at the resonant frequency of the grid tuned circuit. The remedy for such parasitic oscillations is to insert either a resistance in series with the grid or a choke coil in series with the plate, as shown in Fig. 194*a*. Such a resistance is directly in series with the grid tuned circuit, and so reduces the tendency to oscillate at the parasitic frequency. The plate choke reduces the resonant frequency of the plate tuned circuit, causing it to offer a capacitive reactance at the resonant frequency of

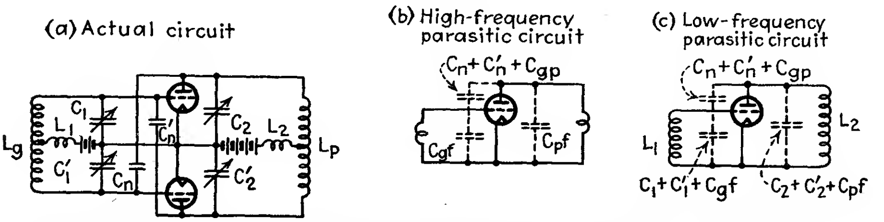


FIG. 195.—Neutralized push-pull amplifier, showing how the tubes can operate in parallel to produce a high-frequency parasite and also a low-frequency parasite.

the grid tuned circuit, and thereby produces a non-oscillatory condition. Neither grid resistance nor plate choke have appreciable effect on the desired mode of operation because they are not in series with the resonant circuits effective for normal operation.

Next consider the push-pull circuit of Fig. 195*a*. An arrangement of this sort is commonly troubled with a high-frequency parasitic oscillation. At this high frequency the capacities C_1, C_1', C_2, and C_2' are effectively short circuits, which places the two tubes in parallel and leads to the equivalent parasitic circuit of Fig. 195*b*, which is a tuned-plate

tuned-grid arrangement, with the neutralizing capacity effectively in parallel with the grid-plate tube capacity and so increasing the tendency to oscillate. The inductance effective in the grid tuned circuit is the inductance of the lead from grid to ground through the capacities C_1 or C_1'. The plate-circuit inductance is similarly the lead inductance, while the tuning capacities are supplied by the tube. This type of parasitic oscillation can be eliminated by inserting resistances in the grid circuits next to the grids or by detuning the plate parasitic circuit by the use of choke coils inserted in the plate leads next to the plates.

The circuit shown in Fig. 195*a* may also develop low-frequency parasitic oscillations in which the grid and plate chokes participate. At these low frequencies the grid and plate tuning inductances L_g and L_p can be considered as short circuits, making the equivalent circuit have the form shown at Fig. 195*c*, which is seen again to be a tuned-plate tuned-grid circuit, with the grid and plate chokes now supplying the tuning inductances. Since these chokes are large and are tuned by a considerable capacity, the frequency of the corresponding parasitic oscillation is commonly quite low. The remedy consists either in arranging the circuit proportions so that the plate tuned circuit is resonant to a lower frequency than the grid tuned circuit, or, better yet, in eliminating one of the chokes.

When tubes are connected in parallel, one can almost always expect parasitic oscillations in which the two tubes operate in series. Thus in Fig. 196*a* one finds that at a very high frequency the normal tuning capacities are essentially short circuits; but the lead inductances from grid to grid and plate to plate in conjunction with the interelectrode capacities of the plate form a tuned-grid tuned-plate circuit, as illustrated in Fig. 196*b*, which may produce parasitic oscillations unless grid resistors or detuning plate inductances are employed. The neutralization has no effect because to the parasitic oscillation the neutralizing condenser is between points that are effectively at ground potential.

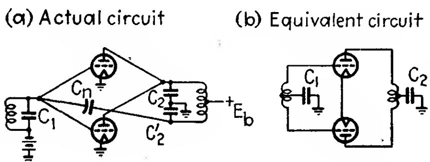


FIG. 196.—Amplifier with parallel tubes, showing how the tubes may act in push-pull to produce a high-frequency parasitic oscillation.

When a tube is operated so that there is sufficient secondary electron emission at the grid to produce a negative resistance, high-frequency oscillations of the dynatron type (see next section) may exist in the grid circuit near the crest of each cycle of the exciting voltage even when tuned-plate tuned-grid oscillations have been eliminated. These oscillations are eliminated in large water-cooled tubes by a grid rectifier such as is illustrated in Fig. 188, which neutralizes the negative resistance;

they can usually be avoided in smaller tubes by slight changes in operating conditions.

The above examples are merely typical of what may be expected. In actual practice it will be found that every new design of equipment has its own peculiarities and its own special varieties of parasitic oscillations, and the number of variations possible appears to be almost without limit. Thus, as an illustration of the complexity of the problem, it is sufficient to note that, in the first 500-kw broadcast transmitters developed, the final power amplifier developed approximately twelve varieties of parasitic oscillations ranging all the way from audible to ultra-high frequencies.

Miscellaneous Comments on Methods of Investigating and Eliminating Parasitic Oscillations.—Parasitic oscillations can be expected as a matter of course in any new design of power amplifier or power oscillator involving large tubes. The simplest method of investigating the presence of parasitic oscillations in a Class C amplifier is to remove the exciting voltage, make the grid bias small or even zero, and operate with a plate voltage lowered sufficiently to keep within the rated dissipation of the tube. This insures a high mutual conductance and gives conditions favorable for the excitation of most types of parasites. These oscillations can then be searched for by means of a neon lamp on a stick. When the lamp is brought near a point of high voltage, it will glow, thus making it possible to determine what parts of the circuit are involved in the parasitic oscillation. After a parasitic oscillation has been thus located and its frequency determined, the equivalent circuit can be deduced and remedial means devised. It is commonly found that, upon the elimination of one parasitic oscillation, another oscillation of a different type will appear, and that upon the elimination of this still other parasitic oscillations may start up.

Power oscillators are less troubled by parasitic oscillations than are Class C amplifiers because the desired oscillation tends to suppress parasitics. Nevertheless, high-frequency parasitic oscillations involving the inductance of the grid and plate leads and the interelectrode capacities of the tube can generally be expected with oscillators employing water-cooled and large air-cooled tubes. These can usually be suppressed by a resistance in series with the grid or a choke connected in the plate lead next to the tube.

Screen-grid and pentode power amplifiers are practically immune from most types of parasitic oscillations provided the screen grid is really at ground potential, since then no energy transfer through the tube is possible. However, the circuit from the screen-grid electrode through the by-pass condenser to cathode has inductance, and, if this circuit is not made as short as possible, the voltage drop in this inductance

will be sufficient to prevent the screen from being at the cathode potential. Feedback through the tube with the possibility of parasitic action will then be present.

Troubles from parasitic oscillations can be minimized by using simple circuits arranged with the shortest possible leads. Neutralization systems that are symmetrical and so maintain the neutralization over a wide frequency band are also helpful. It is also desirable to employ inductive coupling to the input and output circuits of the tube rather than capacitive or tapped inductance coupling. Shunt-feed chokes should be avoided when feasible, and should not be employed in both grid and plate circuits of the same tube. The remedies most commonly successful in suppressing parasitic oscillations are the use of a resistance in series with the grid lead of the tube, the use of a small inductance coil in the plate lead next to the tube, and the use of as few radio-frequency chokes as possible. When these methods fail, it is necessary to make a detailed study of the situation and to devise arrangements to satisfy the circumstances.

69. Miscellaneous. *Synchronization of Vacuum-tube Oscillators.*—Vacuum-tube oscillators have an inherent tendency to synchronize with any other oscillation of approximately the same frequency that may be present. The behavior of two oscillators loosely coupled together and generating frequencies that are not widely different illustrates what can be expected. If the two frequencies differ by only a small percentage, they are both shifted from their normal values in such a way as to reduce the difference. This attraction of the two frequencies becomes more pronounced as the difference between the normal oscillating frequencies is reduced and finally becomes so great that the oscillators pull into synchronism. The extent to which the frequency of an oscillator can be shifted from its normal value by the presence of currents of a slightly different frequency will be greater as the strength of the injected currents is increased and as the frequency stability of the oscillator is lowered.[1]

It has also been found that, when currents having a frequency approximating a harmonic of the generated frequency are injected into the oscillator circuits, there is a tendency for the generated oscillation to synchronize its frequency with the injected currents in such a way that a harmonic of the generated frequency is in exact synchronism with the injected frequency. The tendency toward synchronization with a harmonic of the generated oscillation is not so great as when both frequencies are approximately the same. However, when the stability of the oscillator whose frequency is being controlled is relatively low

[1] A mathematical analysis of this phenomenon of synchronization for one particular type of oscillating circuit is given in E. V. Appleton, Automatic Synchronization of Triode Oscillators, *Proc. Cambridge Phil. Soc.*, vol. 21, p. 231, 1922.

and when the injected energy is relatively strong, it is not difficult to maintain the oscillations at a frequency that is exactly one-half or one-third of the injected frequency, and synchronization has been obtained when the ratio of frequencies is as low as one-sixth.[1]

Oscillators with More Than One Tube.—Since vacuum-tube oscillators tend to synchronize automatically, there is no difficulty in operating two or more tubes in parallel, and such arrangements can be used where more power is desired than can be obtained from a single tube. However, it is usually preferable to use a single tube of larger capacity where this is possible, because parasitic oscillations are very difficult to avoid when tubes are operated in parallel.

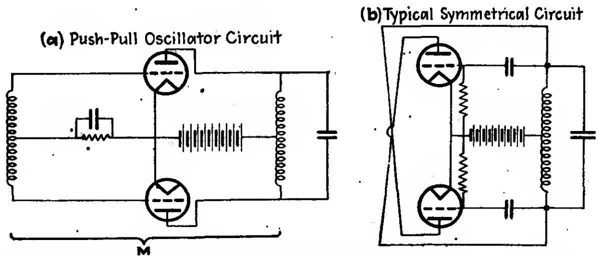


FIG. 197.—Typical symmetrical two-tube oscillator circuits. (*a*) is a push-pull circuit of the reversed feedback type, while (*b*) is a symmetrical oscillator derived from the Hartley circuit.

The most successful multi-tube arrangements are those employing two tubes in a push-pull or some other symmetrical connection, such as shown in Fig. 197, since with such arrangements there is less tendency for parasitic oscillations to be produced. Symmetrically arranged oscillators using two tubes are often employed where the frequency to be obtained is extremely high because they permit the use of very short connecting wires and in effect connect the electrode capacities of the two tubes in series, which increases the highest frequency to which the resonant circuit can be tuned.

Dynatron and Similar Oscillators Employing Negative Resistance.[2]—The negative plate-cathode resistance of a dynatron can be used to produce oscillations by connecting a parallel resonant circuit in series with the negative resistance as shown in Fig. 198*a*. Such an arrangement will oscillate provided the negative plate resistance is less than the

[1] Synchronization between oscillations related by harmonics is discussed in the paper by Isaac Koga, A New Frequency Transformer or Frequency Changer, *Proc. I.R.E.*, vol. 15, p. 669, August, 1927. Also see U. S. Patent 1,527,228 issued to Schelleng; and J. Groszkowski, Frequency Division, *Proc. I.R.E.*, vol. 18, p. 1960, November, 1930.

[2] For further discussion see F. E. Terman, "Measurements in Radio Engineering," 1st ed., pp. 287–290, McGraw-Hill Book Company, Inc.

parallel resonant impedance of the tuned circuit. Also, if the grid bias is such that the negative resistance is just barely low enough for oscillation, the frequency stability with respect to tube conditions is extremely high and the oscillation is practically sinusoidal. Such oscillators are sometimes used for laboratory or test oscillators, and are particularly satisfactory when combined with automatic amplitude control so as to maintain the tube conditions on the threshold of oscillation at all times, as discussed below.

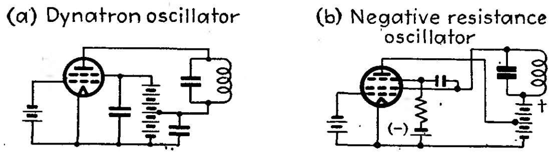


FIG. 198.—Oscillators employing negative resistance supplied by a tube.

Negative resistances obtained in other ways can be used in similar manner to produce oscillations. Thus the oscillator of Fig. 198*b* employs the negative resistance arrangement of Fig. 83, and has the advantage over the dynatron in that negative resistance obtained by secondary emission is relatively unstable at times.

Oscillators with Automatic Amplitude Control.[1]—The frequency stability and wave-shape form of any common oscillator can be improved

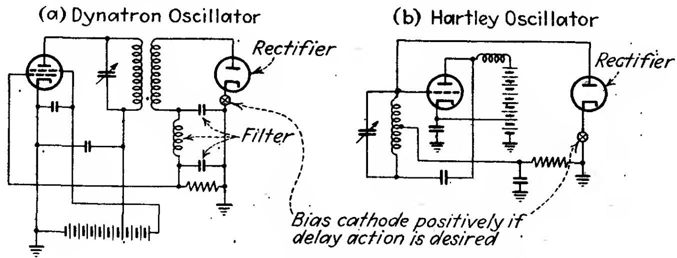


FIG. 199.—Oscillators with simple amplitude control.

by using an automatic-amplitude-control arrangement to maintain the amplitude of oscillations constant under all conditions. This is particularly the case when the control is adjusted so that the oscillation is restricted to the linear part of the tube characteristic.

The amplitude control commonly takes the form of an automatic-volume-control circuit which may or may not have delay action. Such

[1] For further information on such oscillators see *ibid.*, pp. 288–291. Also see L. B. Arguimbau, An Oscillator Having a Linear Operating Characteristic, *Proc. I.R.E.*, vol. 21, p. 14, January, 1933; Janusz Groszkowski, Oscillators with Automatic Control of the Threshold of Regenerations, *Proc. I.R.E.*, vol. 22, p. 145, February, 1934.

circuit arrangements applied to a dynatron and a conventional Hartley oscillator are illustrated in Fig. 199.

Resistance-stabilized Oscillators.[1]—The resistance-stabilized oscillator is widely used to generate audio and low radio frequencies for laboratory and similar uses where plate efficiency can be sacrificed to frequency stability and good wave form. Typical circuit arrangements are shown in Fig. 200 and are seen to be conventional oscillator circuits with the addition of a "feedback" resistance located between the plate of the oscillator tube and the tuned circuit. This feedback resistance must be high compared with the plate resistance of the tube, and it has two primary functions. First, it makes the resistance that the tuned circuit sees when looking toward the plate independent of the electrode voltages;

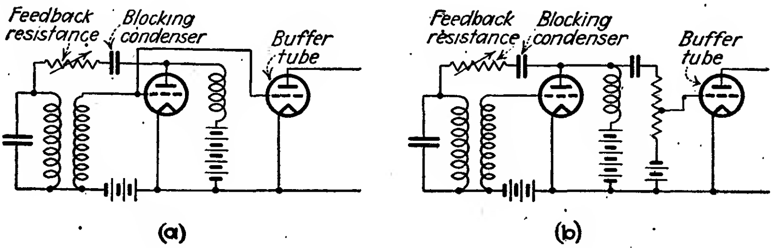


FIG. 200.—Typical oscillator circuits provided with resistance stabilization.

and, second, it provides a means for limiting the amplitude of oscillations to the straight-line part of the tube characteristic.

The tube in a resistance-stabilized oscillator is adjusted to operate as a Class A amplifier, and the feedback resistance is made so high that oscillations are just barely able to start. Under these conditions oscillations build up until there is grid current, which introduces additional losses that increase rapidly with further increase in amplitude. If the feedback resistance is so high that oscillations are barely able to exist with no grid loss, an equilibrium will be reached at an amplitude that drives the grid only a few volts positive. It will be noted that a fixed grid bias, such as is obtained from a biasing resistance, is necessary and that the grid-leak bias arrangement commonly used with power oscillators is not permissible.

The wave form is determined by the linearity of the tube's dynamic characteristic over the range of voltage that the oscillations apply to the grid. It is apparent that for good wave form, the tube, when considered as an amplifier, must be so adjusted that it will amplify without distortion an alternating-current voltage on the grid having a crest value slightly

[1] For a more detailed discussion of this type of oscillation see *ibid.*, pp. 283–287, or F. E. Terman, Resistance-stabilized Oscillators, *Electronics*, vol. 6, p. 190, July, 1933.

greater than the grid bias. This means that the oscillator tube should be operated at a grid bias which is slightly less than the bias that would be used for Class A amplifier operation at the same plate voltage and without grid current.

Best results are obtained when attention is paid to certain circuit details. The circuit proportions should be such that the feedback resist-

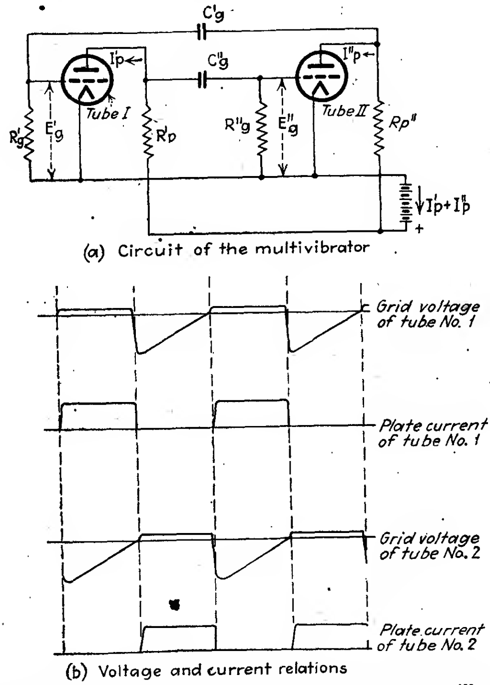


FIG. 201.—Circuit of the multivibrator, together with oscillograms, showing the way in which the instantaneous grid potential and plate currents vary during the cycle of operation.

ance required is at least twice, and preferably over five times, the plate resistance. The blocking condenser in series with the feedback resistance must have a low reactance compared with this resistance in order to avoid phase shifts, while the shunt-feed choke should have a reactance that is high compared with the plate resistance of the tube for the same reason. The frequency stability is also helped greatly by making the coupling between plate and grid coils as close as possible.

The Multivibrator (or Relaxation Oscillator).[1]—The multivibrator is a two-stage resistance-coupled amplifier in which the voltage developed

[1] For a more detailed discussion of the multivibrator see F. E. Terman, "Measurements in Radio Engineering," 1st ed., pp. 129–136, McGraw-Hill Book Company, Inc.

by the output of the second tube is applied to the input of the first tube as shown in Fig. 201. Such an arrangement will oscillate because each tube produces a phase shift of 180°, thereby causing the output of the second tube to supply to the first tube an input voltage that has exactly the right phase to sustain oscillations. The usefulness of the multivibrator arises from the fact that the wave that is generated is very rich in harmonics and from the fact that the frequency of oscillation is readily controlled by an injected voltage.

The operation of the multivibrator can be understood by reference to the oscillograms shown in Fig. 201*b*. Oscillations are started by a minute voltage at the grid of one of the tubes, say a positive potential on the grid of Tube I. This voltage is amplified by the two tubes and reappears at the grid of the first tube to be reamplified. This action takes place almost instantly and is repeated over and over, so that the grid potential of Tube I rises suddenly to a positive value, while the grid potential of Tube II just as suddenly becomes more negative than cut-off. The immediate result is that amplification ceases, and for the moment one tube is drawing a heavy plate current while the other tube takes no plate current. This situation is not permanent, however, because the leakage through the grid-leak resistances gradually brings the grid potentials back toward normal. When this leakage has reached the point where amplification is just on the verge of being possible, some minute voltage will change the potentials enough to start the amplification process in the reverse direction, *i.e.*, the grid of Tube I will suddenly become negative and the grid of Tube II positive. This action is clearly evident in the oscillograms and is exactly the same as the initial action except that the relative functions of the two tubes have been interchanged. Next the potentials on the two grids gradually die away as the result of the action of the grid leaks, just as before, and finally reach a point at which the cycle repeats.

The frequency of the multivibrator oscillation is determined primarily by the grid-leak resistance and grid-condenser capacity, but it is also influenced by the remaining circuit constants, the tube characteristics, and the electrode voltages. The multivibrator can be adjusted to generate frequencies ranging anywhere from perhaps 1 cycle per minute to frequencies in excess of 100,000 cycles per second. The upper limit is the highest frequency at which satisfactory resistance-coupled amplification is possible, while the lower limit is fixed by the leakage of the grid condenser in relation to the condenser capacity.

Injection of an alternating voltage into the multivibrator circuit tends to cause the latter to adjust itself to a frequency that is m/n of the injected frequency, where m and n are integers. It is possible in this way to use the multivibrator to reduce frequency (*i.e.*, to generate a subharmonic

of the injected frequency), and it is entirely practicable to maintain the control rigidly when n is as large as 10 and m is 1.

The principal use of the multivibrator is in the measurement of frequency. The multivibrator oscillations have a wave that is rich in harmonics, so by using a standard frequency of known value to control the frequency of the multivibrator it is possible to obtain many frequencies related to the standard. If the fundamental frequency of the multivibrator is controlled by the fundamental frequency or a harmonic of the standard, the multivibrator will then produce higher harmonics of the standard frequency. On the other hand, when the frequency of the multivibrator oscillations is controlled in such a way as to be an exact subharmonic of the standard frequency, the multivibrator produces frequencies that are less than the standard. Thus, when the multivibrator oscillation has a fundamental frequency that is exactly one-tenth of the standard controlling frequency, the harmonics of the multivibrator are exactly 1/10, 2/10, 3/10, etc., of the standard frequency, and the arrangement is in effect a frequency-reducing system.

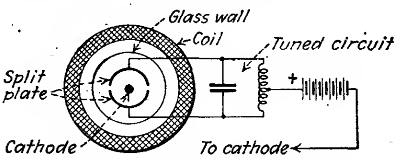


FIG. 202.—Split-anode type of magnetron-oscillator circuit. When the field strength is just greater than the critical value, oscillations are developed which have the frequency of the resonant circuit.

The Magnetron.—Magnetron tubes can be used to generate oscillations in several ways. One method is to utilize the magnetic field to turn the anode current on and off.[1] A more common arrangement utilizes the negative resistance characteristic between the two anodes of a split-anode magnetron, which was discussed in Sec. 36, and employs the circuit shown in Fig. 202. Such split-anode oscillators have been used to a limited extent at ultra-high frequencies between the frequency region at which ordinary triodes function satisfactorily and frequencies which are so high that electron oscillators must be employed.[2] Magnetrons are also used to produce electron oscillations as discussed in Sec. 71.

70. Crystal Oscillator.[3]—The frequency stability of an oscillator can be made very high by replacing the usual resonant circuit with a mechan-

[1] Such an oscillator is described by Frank R. Elder, The Magnetron Amplifier and Power Oscillator, *Proc. I.R.E.*, vol. 13, p. 159, April, 1925.

[2] See W. C. White, Producing Very High Frequencies by Means of the Magnetron, *Electronics*, vol. 1, p. 34, April, 1930.

[3] The literature on crystal oscillators and related piezo-electric phenomena is too extensive to be covered with any thoroughness in a book of this type. The following selected bibliography, in addition to the footnote references in this section, is recommended to the reader desiring additional background: W. G. Cady, The Piezo-electric Resonator, *Proc. I.R.E.*, vol. 10, p. 83, April, 1922; Bibliography on Piezo-electricity, *Proc. I.R.E.*, vol. 16, p. 521, April, 1928; Electroelastic and Pyro-electric Phenomena,

ically vibrating piezo-electric quartz crystal and utilizing the piezo-electric effect to obtain the connection between the electrical circuits and the mechanical vibrations. The piezo-electric material used in crystal oscillators comes in the form of crystals which, when complete, have a hexagonal cross section and pointed ends, as illustrated in Fig. 203.[1] The properties of such a crystal can be expressed in terms of three sets of axes. The axis joining the points at the ends of the crystal is known as the optical axis, and electrical stresses applied in this direction produce no piezo-electric effect. The three axes X', X'', and X''' passing through the corners of the hexagon that forms the section perpendicular to the optical axis are known as the electrical axes, while the three axes Y', Y'', and Y''', which are perpendicular to the faces of the crystal, are the mechanical axes.

If a flat section is cut from a quartz crystal in such a way that the flat sides are perpendicular to an electrical axis as indicated in Fig. 204*b* (X or Curie cut), it is found that mechanical stresses along the Y-axis of such a section produce electrical charges on the flat sides of the crystal section. If the direction of these stresses is changed from tension to compression or vice versa, the polarity of the charges on the crystal surfaces is reversed. Conversely, if electrical charges are placed on the flat sides of the crystal by applying a voltage across these faces, a mechanical stress is produced in the direction of the Y-axis. This property by which mechanical and electrical properties are interconnected in a crystal is known as the piezo-electric effect and is exhibited by all sections cut from a piezo-electric crystal. Thus, if mechanical forces are applied across the faces of a crystal section having its flat sides perpendicular to a Y-axis, as in Fig. 204*a* (which is known as the Y or 30° cut), piezo-electric charges will be developed because forces and potentials developed in such a crystal have components across the Y- and X-axes, respectively.

Proc. I.R.E., vol. 18, p. 1247, July, 1930; Piezo-electric Terminology, *Proc. I.R.E.*, vol. 18, p. 2136, December, 1930; A. Crossley, Piezo-electric Crystal-controlled Transmitters, *Proc. I.R.E.*, vol. 15, p. 9, January, 1927; August Hund, Uses and Possibilities of Piezo-electric Oscillators, *Proc. I.R.E.*, vol. 14, p. 447, August, 1926; Note on Quartz Plates, Air-gap Effect, and Audio-frequency Generation, *Proc. I.R.E.*, vol. 16, p. 1072, August, 1928; Summary of Piezo-electric Crystal Conference Held by U. S. Navy Department, December 3–4, 1929, *Proc. I.R.E.*, vol. 18, p. 2128, December, 1930. Further information on the practical side of crystal cutting and grinding is to be found in "Radio Amateurs' Handbook," published by the American Radio Relay League, and in "The Radio Handbook," Pacific Radio Publishing Company. Crystal oscillators are the invention of Dr. W. G. Cady.

[1] There are many crystalline substances which have piezo-electric properties, such as Rochelle salts, tourmaline, quartz, etc., but of these quartz is used exclusively in crystal oscillators because of its cheapness, mechanical ruggedness, and low temperature coefficient.

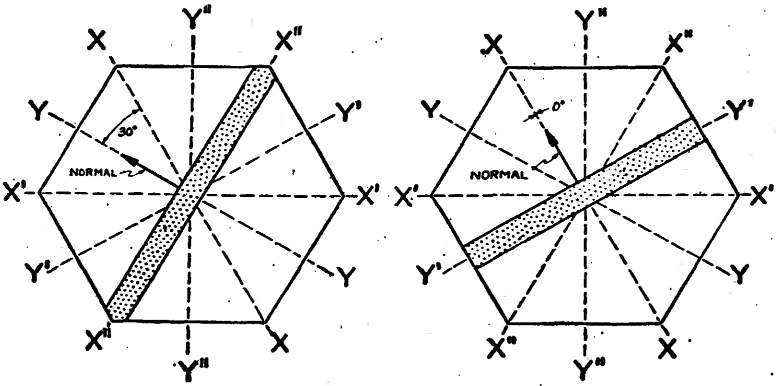
When an alternating voltage is applied across a quartz crystal in such a direction that there is a component of electric stress in the direction of an electrical axis, alternating mechanical stresses will be produced in the direction of the Y- (or mechanical) axis which is perpendicular to the X-axis involved. These stresses will cause the crystal to vibrate, and, if the frequency of the applied alternating voltage approximates a frequency at which mechanical resonance can exist in the crystal, the amplitude of the vibrations will be very large. In the vicinity of such a resonant frequency the current that is drawn by the crystal as a result of the vibrations is exactly the same current that would be drawn by a series circuit composed of resistance, inductance, and capacity. In addition to this current representing the vibrational characteristics of the piezo-electric crystal, there is also a component of leading current resulting from the electrostatic capacity between the points of application of the exciting voltage.

FIG. 203.—Illustration showing the natural quartz crystal and the relation of the electric or X-, the mechanical or Y-, and the optical or Z-axes to the crystal structure. The upper section shows a Y (or 30°) cut plate while the plate in the center section is X (or Curie) cut. The third Y-axis $Y''Y''$ is not shown because the perspective of the drawing makes it coincide with the ZZ-axis.

Equivalent Electrical Circuit of Quartz Crystal.—As far as the electrical circuits associated with the vibrating crystal are concerned, the crystal can be replaced by the electrical network of Fig. 205 in which C_1 represents the electrostatic capacity between the crystal electrodes when the crystal is not vibrating, and the series combination L, C, and R represents the electrical equivalent of the vibrational characteristics of the material.[1] The inductance L is the electrical equivalent of the crystal mass that is effective in the vibration, C is the electrical equivalent of the effective mechanical compliance, while R represents the electrical equivalent of the coefficient of friction. The frequency at which L and C are in series resonance is also the frequency of mechanical resonance. The electrical energy drawn by the equivalent L-C-R series

[1] The fact that a vibrating quartz crystal can be replaced by an equivalent electrical network was first discovered by Dr. H. J. Ryan and his co-workers during war researches on submarine detection carried on in 1918, but the report representing this work has not yet been released by the military authorities. The equivalent electrical circuit of the vibrating piezo-electric crystal has since been discovered by several other investigators. For further information see K. S. Van Dyke, The Piezo-electric Resonator and Its Equivalent Network, *Proc. I.R.E.*, vol. 16, p. 742, June, 1928; W. P. Mason, An Electromechanical Representation of a Piezo-electric Crystal Used as a Transducer, *Proc. I.R.E.*, vol. 23, p. 1252, October, 1935.

circuit represents energy that the electrical circuit supplies to maintain the crystal vibrations. Below resonance this energy contains a leading



a b

FIG. 204.—Cross sections of the quartz crystal shown in Fig. 203 taken in planes taken perpendicular to the optical axis ZZ. The plate at a has the Y cut (also called 30° cut) because its face is perpendicular to a Y- (*i.e.*, mechanical) axis, while the plate at b is an X-cut plate (Curie cut) because its face is perpendicular to an X- (*i.e.*, electric) axis.

reactive component because the elastic forces control the crystal vibration, while above resonance the inertia is the dominating factor and lagging reactive energy is required to sustain vibrations. At resonance the vibrations consume no reactive energy, and the crystal consumes power at unity power factor. Electrical circuits involving piezo-electric crystals can therefore be analyzed by replacing the crystal with its equivalent electrical network and then determining the behavior of the resulting circuit.

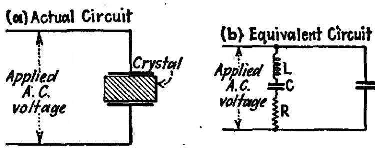


FIG. 205.—Equivalent electrical network that represents the effect which a vibrating quartz crystal has on the electrical circuits associated with it.

The magnitudes of L, C, R and C_1 that enter into the equivalent electrical network of the vibrating quartz crystal depend upon the way in which the crystal is cut, the size of the crystal, and the type of vibration involved. The numerical values can be calculated from the crystal dimensions for the simpler modes of vibrations, and typical values for several crystals are given in Table X.[1]

[1] In the case of X- (Curie-) cut crystals in which the principal vibration is in the direction of the X-axis (thickness vibration) the electrical quantities are given to a fair approximation by the following formulas:

$$L_t = 130\frac{t^3}{lw} \text{ henries} \tag{177a}$$

$$C = 0.0022\frac{lw}{t}\ \mu\mu f \tag{177b}$$

$$C_1 = 0.40\frac{lw}{t}\ \mu\mu f \tag{177c}$$

When the principal vibration is in the direction of the Y-axis (width vibration), the

TABLE X.—EQUIVALENT ELECTRICAL CHARACTERISTICS OF TYPICAL QUARTZ CRYSTALS

Crystal No.	Dimensions, cm			Type of cut	Type of vibration	Resonant frequency, kc	Equivalent electrical quantities				
	t	w	l				L, henries	C, $\mu\mu$f	R, ohms	C_1, $\mu\mu$f	Q, approximately
I	0.15	3.0	0.40	X	Width	About 90	137	0.0235	About 7,500	3.54	10,300
II	0.25	2.5	2.5	X	Thickness	About 1,100	0.33	0.065	About 2,700	1.0	844
III	0.636	3.33	2.75	X	Thickness	451.5	3.656	0.0316	9,036	5.755	1,147

The frequency at which mechanical resonance takes place is the frequency at which L and C are in resonance, and the magnitude of the resonance effect is determined by the ratio $\omega L/R$ of the equivalent electrical network (*i.e.*, by the equivalent Q of the crystal). The outstanding characteristics of the crystal vibrator are that the resonant frequency varies inversely with the dimensions of the crystal in the direction in which the principal vibration is taking place, that the ratio L/C of the equivalent quartz resonator is enormously higher than could conceivably be obtained with coils and condensers, and finally that the effective Q of the crystal vibrator is extremely high, particularly when the frequency of vibration is low.

Oscillating Crystals.—Since the vibrating quartz crystal is equivalent to a resonant circuit, it can be used as the frequency-controlling element in a vacuum-tube oscillator in place of the usual tuned circuit. While many circuit arrangements can be employed to accomplish this, the one used in most commercial equipment is shown in Fig. 206. When the crystal is replaced by its equivalent electrical circuit, this is seen to be a tuned-plate tuned-grid oscillator circuit, in which the grid tuned circuit

---

formulas for C, C_1, and R are the same, while the equivalent inductance L becomes

$$L\omega = 130\frac{wt}{l} \tag{177d}$$

The dimensions w, l, and t are measured in centimeters in the direction of the Y-axis parallel to the surface of the crystal, in the direction of the X-axis perpendicular to this Y-axis, and in the direction of the Z-axis, respectively.

The effective resistance of the crystal depends largely upon the method of mounting, and the Q is often much higher than the values in the Table. Karl S. Van Dyke, in A Determination of Some of the Properties of the Piezo-electric Quartz Resonator, *Proc. I.R.E.*, vol. 23, p. 386, April, 1935, finds that a particular crystal, which had a Q of 25,000 when mounted in air, had a Q of 580,000 when etched to remove grinding particles and mounted in a vacuum to eliminate radiation losses.

is supplied by the crystal. The coupling between the tube and this circuit (and hence the crystal) is determined by the ratio C/C_1, and by reference to Table X is seen to be small. This small coupling, combined with the extremely high Q of the crystal vibrator and the permanence of the quartz plate, makes the frequency stability of the crystal oscillator very high.

Crystal Cuts and Modes of Oscillation.—The crystals used in crystal oscillators are commonly in the form of plates cut from the natural crystal with orientations known as X, Y, or AT cuts.

The relationship between the X-cut plate and the crystal axis is shown in Fig. 203. Such a plate has two principal modes of oscillation.

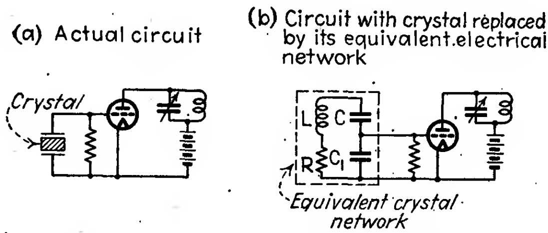


FIG. 206.—Circuit of crystal oscillator, together with equivalent circuit in which the crystal is represented by an electrical network.

One of these is determined by the thickness of the plate in the direction of the electrical axis and is given to a good approximation by the equation

$$\left.\begin{array}{l}\text{Thickness frequency}\\ \text{of } X\text{-cut plate}\end{array}\right\} = \frac{2.86 \times 10^6}{t} \tag{178}$$

where t is the thickness in millimeters as measured in the direction of the X-axis perpendicular to the crystal surface. The other frequency is determined by the width of the crystal in the direction of the Y- or mechanical axis, and it is given by Eq. (178) by substituting for t the width w of the crystal in millimeters measured in the direction of the Y-axis. These two frequencies of oscillation correspond to standing waves along the thickness and width dimensions of the crystal plate, respectively, and Eq. (178) can be predicted from the appropriate elastic constants for quartz. The temperature coefficient of both these resonant frequencies is negative and has a value of about 20 parts in a million per degree centigrade. The circuit of Fig. 206 can be made to excite either the width or thickness vibration at will by adjusting the plate tuned circuit to be resonant at a frequency slightly higher than the desired vibration.

The relationship between a Y-cut plate and the crystal axes is shown in Fig. 203. Such a crystal has two principal resonant frequencies corresponding to those obtained with the X cut. The width frequency

can be obtained from Eq. (178) by substituting for the quantity t appearing in the equation the width w corresponding to the width along the electrical axis parallel to the surface of the crystal. The thickness frequency is given approximately by the following equation

$$\left.\begin{array}{l}\text{Thickness frequency in}\\ Y\text{-cut crystal}\end{array}\right\} = \frac{1.96 \times 10^6}{t} \tag{179}$$

where t is the thickness in millimeters in the direction of the Y-axis perpendicular to the crystal surface. The thickness frequency is some-

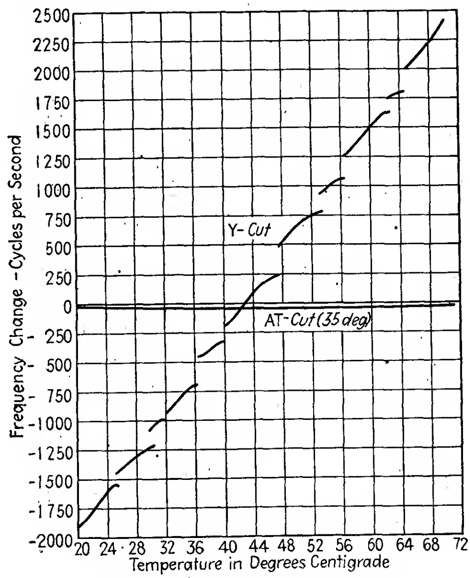


FIG. 207.—Temperature-frequency characteristic curve of Y- and AT-cut plates showing superiority of latter with respect to low-temperature coefficient and continuity of characteristic. The crystal frequency is 1600 kc.

what dependent upon the width of the crystal and also exhibits other peculiarities as discussed below. The temperature coefficient of the width vibration of a Y-cut crystal is approximately 20 parts in a million per degree centigrade and is negative, while the thickness vibration has a temperature coefficient that may vary from -20 to $+100$ parts in a million per degree centigrade, with the exact value depending upon the operating temperature and ratio of width to thickness.

A quartz crystal represents a very complex vibrating system having many degrees of freedom. Thus it is possible to obtain harmonics of the principal modes of vibration discussed above, as well as flexural and torsional vibrations, etc.[1] Many of these modes of vibration are mechan-

[1] For discussions of such vibrations and for means of exciting them, see J. R. Harrison, Piezo-electric Resonance and Oscillatory Phenomena with Flexural Vibra-

ically coupled to each other in the crystal, with the result that, when a crystal is operated in such a way as to produce one of the principal modes of oscillation, other modes of oscillation are simultaneously excited. This is particularly true of Y-cut crystals, and it causes thin Y-cut plates excited at the thickness frequency to have several frequencies of oscillation usually quite close together. Such Y-cut plates also commonly have a discontinuous temperature-frequency characteristic as shown in Fig. 207, as well as a discontinuous thickness-frequency curve and a thickness frequency that depends somewhat upon the width. This behavior comes about through coupling between the normal thickness mode of oscillation and harmonics of lower frequency oscillations, particularly the width oscillation.

The equivalent circuit taking into account this situation is therefore as shown in Fig. 208, where ω_0 represents the equivalent resonant circuit for the thickness vibration and ω_1, ω_2, etc., represent various other vibrations that are coupled mechanically to the thickness vibration. When a coupled subsidiary vibration has a resonant frequency approximately the same as the primary mode of oscillation, the resonant frequency of the combination as viewed from the crystal electrodes (terminals AA in Fig. 208) is affected appreciably by the coupled reactance and resistance. When the coupling is close, two resonant frequencies differing by only a few kilocycles can be expected, and under some conditions the resistance that the subsidiary vibration couples into the circuit is sufficient to prevent the crystal from oscillating. Inasmuch as the subsidiary vibrations depend in most cases upon the width of the crystal, the exact frequency of oscillation can be modified by edge grinding, and edge grinding will also usually make a non-oscillatory crystal operate. The subsidiary vibrations have a different temperature coefficient from the primary mode of vibration, causing the combination to exhibit the peculiar temperature-frequency characteristic shown in Fig. 207, with discontinuities appearing each time a subsidiary mode of oscillation becomes resonant at the same frequency as the primary mode.

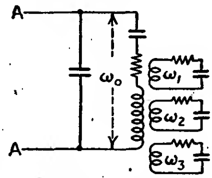


FIG. 208.—Equivalent electrical circuit of crystal, showing the subsidiary resonances ω_1, ω_2, ω_3, etc., which can be considered as coupled to the main resonance ω_0.

The subsidiary resonances are more numerous the thinner the crystal, and they cause very thin crystals to have a complicated frequency

tions in Quartz Plates, *Proc. I.R.E.*, vol. 15, p. 1040, December, 1927; August Hund and R. B. Wright, New Piezo-electric Oscillators with Quartz Cylinders Cut along the Optical Axis, *Proc. I.R.E.*, vol. 18, p. 741, May, 1930; N. H. Williams, Modes of Vibration of Piezo-electric Crystals, *Proc. I.R.E.*, vol. 21, p. 990, July, 1933.

spectrum such as illustrated in Fig. 209. The increasing complexity of this frequency spectrum as the plate becomes thinner sets a practical limit to the highest frequency that it is feasible to generate in a crystal oscillator using X- and Y-cut crystal plates.[1]

The couplings between the thickness vibration and other modes of oscillation can be controlled by cutting the plate out of a plane that is rotated about the X-axis as shown in Fig. 210. By making the angle θ have a value of 31°, the coupling between the width and thickness vibrations can be reduced to zero. This simplifies the frequency spec-

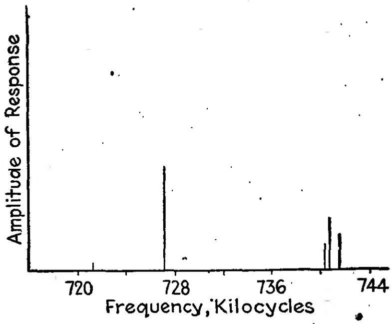


FIG. 209.—Typical frequency spectrum of thin quartz plate, showing how several closely spaced resonant frequencies can be expected. The lengths of the vertical lines indicate in a rough way the relative tendency to oscillate at the different resonant frequencies.

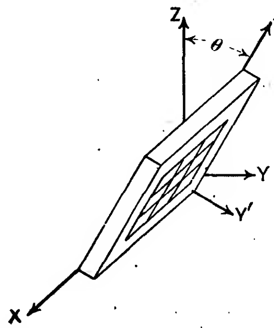


FIG. 210.—Diagram showing how crystal plane is rotated about X-axis to obtain the AT cut.

trum and eliminates practically all the discontinuities in the temperature-frequency and thickness-frequency characteristics. The temperature coefficient depends upon the angle θ; if $\theta = 35°$, the crystal will have substantially zero change in resonant frequency with temperature, as illustrated in Fig. 207, and at the same time will have only very small coupling between the width and thickness vibration. A crystal obtained in this way with $\theta = 35°$ is known as the AT cut and is rapidly displacing the X and Y cuts for high-frequency operation.[2]

[1] For further information on the effects produced by subsidiary resonances, see F. R. Lack, Observations on Modes of Vibration and Temperature Coefficients of Quartz Crystal Plates, *Proc. I.R.E.*, vol. 17, p. 1123, July, 1929.

This paper shows how it is possible by the choice of proper ratio of width to thickness in a Y-cut plate to balance the positive temperature coefficient of the thickness vibration against the negative coefficient of the coupled width vibrations to give a resultant zero temperature coefficient over a limited temperature range.

[2] For a more complete discussion concerning the effect of rotating the crystal plane about the X-axis, see F. R. Lack, G. W. Willard, and I. E. Fair, Some Improve-

Relative Merits of Different Cuts.—The AT cut is by far the most satisfactory for frequencies above about 300 kc. It has a resonant frequency that is substantially independent of temperature, so that constant-temperature operation is usually unnecessary. The frequency spectrum is simple, and discontinuous changes of frequency with temperature, width, thickness, etc., are few, so that the manufacturing is relatively simple and very thin crystals corresponding to high frequencies can be used without trouble from multiple resonant frequencies.

At the lower frequencies X- or Y-cut crystals employing the width vibration are used. The temperature coefficient of these cuts is such that, when high frequency stability is required, the crystal must be maintained at constant temperature by means of a suitable thermostatically controlled oven. Until the development of the AT cut, thickness vibrations of X- and Y-cut plates were also universally used for the generation of high frequencies. The X-cut crystals at high frequencies have a simpler frequency spectrum and give less trouble from discontinuities of the frequency than do Y-cut crystals, but the Y-cut crystal has the advantage that the oscillations are of a shear type that permits the edges of the crystal to be clamped by the holder without seriously affecting the vibrations. This is important in the case of apparatus which is portable or which is subject to mechanical vibrations.

Power Obtainable from Crystal Oscillators.—The power obtainable from a crystal oscillator is limited at high frequencies by heating of the crystal and at low frequencies by the strains which the vibrations set up in the crystal structure and which will crack the crystal if the vibrations are too intense. A crystal having a large area will develop more power than a small crystal at the same resonant frequency, but the extent to which one can go in this direction is limited by the difficulty of grinding large crystals absolutely true.

By pushing the crystal to the limit it is possible to obtain sufficient excitation for controlling a tube generating approximately 50 watts of power at high frequencies. However, working the crystal this hard results in appreciable dissipation of power within the crystal, which causes the temperature to rise. Since X- and Y-cut crystals have a high temperature coefficient and also a discontinuous frequency-temperature relationship, it is very desirable to operate such crystals with relatively light load, the common practice being to control only a few watts of power. With the AT cut the operation of the crystal lightly loaded is not so important, since the temperature can be allowed to rise appreciably without change in the frequency.

---

ments in Quartz Crystal Circuit Elements, *Bell System Tech. Jour.*, vol. 13, p. 453, July, 1934.

Crystal Mountings.[1]—The crystal holder must be arranged to add as little damping as possible to the crystal vibrations, and yet it should, when possible, hold the crystal rigidly in position. X- and Y-cut plates in which the width vibration is utilized can be clamped in the middle of the width direction, since this is a nodal point for the vibrations. Y-cut crystals in which the thickness vibration is utilized can be clamped by spring pressure between electrodes, which are preferably cupped so that the pressure is exerted at only a few points, such as the corners of a square crystal or the periphery of circular crystals. Such clamping has little effect on the thickness vibrations of Y-cut crystals; and at the same time it materially simplifies the frequency spectrum by inhibiting certain of the undesired modes of oscillation that give rise to discontinuities in the frequency-temperature curve, etc. With X-cut crystals using the thickness vibration, it is usually desirable to allow an air gap between the upper electrode and the crystal because clamping introduces appreciable damping. This makes X-cut crystals impracticable in equipment that is portable or subject to vibration. AT-cut crystals can be clamped in the same way as Y-cut crystals.

The variation of the resonant frequency with temperature in X- and Y-cut crystals makes it necessary to maintain such crystals at a constant temperature if high frequency stability is to be maintained. With the AT cut the temperature coefficient of frequency is so low that temperature control can normally be dispensed with. The simplest arrangement for maintaining the crystal at constant temperature utilizes a large copper plate for the lower electrode. The bottom side of this plate is provided with a recess in which is mounted the heater and thermostat, and the entire assembly including the upper plate is mounted in a suitable case of bakelite or isolantite. The large copper plate serves to distribute the heat equally and also to equalize for the intermittent operation of the heater. When the requirements for frequency stability are particularly severe, as in the case of standard frequency equipment, more elaborate constant-temperature systems are employed. These commonly involve at least two concentric compartments thermally insulated from each other and surrounded by a heater that is controlled by a thermostat. In some cases a double thermostat and heater combination is employed.[2]

[1] For further discussions of crystal mountings, see O. M. Hovgaard, Application of Quartz Plates to Radio Transmitters, *Proc. I.R.E.*, vol. 20, p. 767, May, 1932; Vincent Heaton and E. G. Lapham, Quartz Plate Mountings and Temperature Control for Piezo Oscillators, *Proc. I.R.E.*, vol. 20, p. 261, February, 1932.

[2] For details of units for maintaining constant crystal temperature, see J. K. Clapp, Temperature Control for Frequency Standards, *Proc. I.R.E.*, vol. 18, p. 2003, December, 1930; W. A. Marrison, Thermostat Design for Frequency Standards, *Proc. I.R.E.*, vol. 16, p. 976; July, 1928.

Frequency Stability of Crystal Oscillators.—The frequency of a crystal oscillator depends primarily upon the resonant frequency of the crystal, but is affected to a small extent by the constants of the electrical circuits. When the crystal is maintained at constant temperature or when a zero-temperature coefficient cut is employed, a stability of 10 to 30 parts in a million can readily be obtained over long periods of time under commercial conditions if the voltages applied to the oscillator tube are reasonably constant. When every possible precaution and refinement is taken, including such measures as maintaining the associated electrical circuits as well as the crystal at constant temperature, frequency stabilities as high as one part in ten million or better can be obtained.[1]

The frequency of a crystal oscillator is determined primarily by the crystal dimensions, so that, in order to change the frequency, it is usually necessary to employ another crystal or to grind the original crystal if the new frequency is higher. Small changes in the frequency can be made without grinding the crystal by varying a capacity shunted across the crystal holder and, in the case of X- and Y-cut crystals, by varying the crystal temperature.

71. Electron Oscillators. *Barkhausen Oscillators.*[2]—Frequencies higher than obtainable by the use of conventional oscillators can be generated by means of electron oscillators, either of the Barkhausen or of the magnetron type. The nature of the Barkhausen electron oscillator can be understood by reference to Fig. 211, which shows a triode operated with the grid at a high positive potential and the plate at a slight negative potential. Electrons emitted from the cathode under these circumstances are attracted toward the grid but most of them pass through the spaces between the grid wires into the grid-plate space, where they slow down and ultimately stop just before reaching the plate. The electrons are then drawn back toward the grid with increasing velocity, but, if they are not captured by the grid wires, they will pass on into the grid-cathode space and slow down upon approaching the filament. This oscillation about the grid may be repeated over and over before the electron is ultimately removed from the tube by a chance impact against

[1] For examples of crystal oscillators with unusually high frequency stability, see L. M. Hull and J. K. Clapp, A Convenient Method for Referring Secondary Frequency Standards to a Standard Time Interval, *Proc. I.R.E.*, vol. 17, p. 252, February, 1929; W. A. Marrison, A High Precision Standard of Frequency, *Proc. I.R.E.*, vol. 17, p. 1103, July, 1929.

[2] The literature on Barkhausen oscillations is very extensive. The reader particularly interested in this subject is referred to the following articles, in addition to those mentioned later in this section: H. E. Hollmann, On the Mechanism of Electron Oscillations in a Triode, *Proc. I.R.E.*, vol. 17, p. 229, February, 1929. H. N. Kozanowski, A New Circuit for the Production of Ultra-short-wave Oscillations, *Proc. I.R.E.*, vol. 20, p. 957, June, 1932.

the grid wires, and represents an electron oscillation having a frequency determined primarily by the tube construction and the grid potential. Several such electron oscillations about the positive grid are illustrated schematically in Fig. 211*b*.

In order for the electron oscillations to develop appreciable power, it is necessary that the electrons vibrate about the grid structure in synchronism. The arrangement illustrated in Fig. 211*a* inherently provides for such synchronization.[1] This can be demonstrated by assuming that there is superimposed upon the grid an alternating voltage having a frequency corresponding approximately to the time taken by an electron to travel from cathode to plate, and by then investigating the interchange of energy between this superimposed voltage and the electrons.

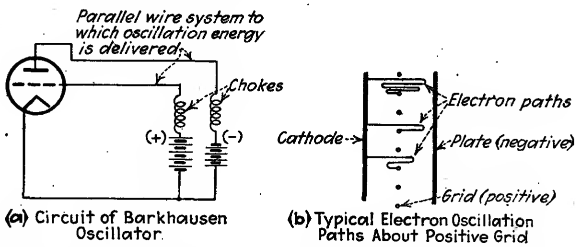


FIG. 211.—Circuit for generating Barkhausen oscillations, together with sketch showing how electron oscillations take place about a positive grid.

Consider first an electron that leaves the cathode at the instant when the superimposed voltage is going through zero and is just starting to become positive. This electron is acted upon by a force consisting of the sum of the direct-current grid potential and the alternating voltage on the grid, and so by the time it reaches the grid plane it is traveling faster than if it had fallen through a potential corresponding to the direct-current grid-cathode voltage. This extra velocity represents energy which the superimposed oscillation has delivered to the electron and which thus tends to damp out the superimposed oscillation. As the electron travels beyond the grid plane into the plate region, it is subjected to braking which retards its motion, but, since the superimposed alternating voltage reverses as the electron passes the grid plane, these braking forces are less than they would be in the absence of a superimposed voltage. The result is that the electron abstracts further energy from the superimposed oscillation. Because of the energy thus obtained from the superimposed oscillation, this electron has sufficient velocity to reach the plate in spite of the slight negative plate potential. This particular electron is therefore of no assistance in maintaining the superimposed

[1] The explanation that follows is due to F. B. Llewellyn, The Barkhausen Oscillator, *Bell Lab. Rec.*, vol. 13, p. 354, August, 1935.

voltage on the grid, but it has the merit of eliminating itself from the picture upon the first opportunity.

Now consider an electron that leaves the cathode one-half cycle later. This electron is attracted to the grid as before, but now the superimposed alternating voltage is of such a polarity during the transit from cathode to grid as to reduce the acceleration that would otherwise be experienced. Furthermore, as the electron passes the grid plane the polarity of the superimposed voltage reverses, so that, as the electron travels toward the plate, it is subjected to increased braking forces and comes to rest before reaching the plate. At the instant the electron starts back toward the grid, the superimposed voltage again reverses, so that the returning electron is subjected to a reduced acceleration until it reaches the grid plane, at which time the superimposed voltage reverses, to cause an increased slowing down of the electron as it travels toward the cathode. The result is that this second electron in its round trip from cathode to plate and back to cathode always works against the superimposed oscillation and therefore delivers to it twice as much energy as the first electron abstracted on its one-way trip. Furthermore, the second or useful electron is still available for future oscillations in which it can give up still more energy. The useful electron continues this process until it ultimately strikes the grid and is lost.

The energy that this electron delivers to the superimposed oscillation is obtained from the plate-supply voltage and represents the difference in the kinetic energy possessed by the electron when it strikes the grid after having been slowed down through several oscillations and the kinetic energy that the electron would have had if it had fallen directly through a potential corresponding to the direct grid-cathode voltage. Electrons leaving the cathode at times intermediate between the departures of the two electrons just considered behave in intermediate fashion. It can be seen that, in a general way, those electrons which abstract energy from the tuned circuit are soon removed, whereas those that give up energy remain and oscillate about the grid a number of times. The result is, therefore, that energy is delivered to the tuned circuit associated with the grid electrode, and with proper adjustment this energy will be sufficient to make the superimposed oscillations maintain themselves.

The above explanation gives a qualitative picture of the mechanism involved in electron oscillators. The actual details are complicated by a number of factors, such as the tendency for space charges to form around the cathode and in the vicinity of the plate, the tendency of the grid to intercept the vibrating electrons, and the fact that, as the electron gives up more and more of its energy, the distance it travels in its vibrations becomes progressively less, so that the electron tends after a number of vibrations to shift its phase with respect to the superimposed oscillation

to the point where it will actually begin to absorb rather than deliver energy.[1] As a result of these conflicting factors, a complete analysis of the electron oscillator has not yet been made.

The frequency of the electron oscillation is determined primarily by the dimensions of the tube and the grid potential, but the external tuned circuit, space charges, negative potential of the plate, etc., may alter the frequency by perhaps 30 to 50 per cent. With cylindrical electrodes and equal grid-cathode and grid-plate spacings, the wave length is given approximately by the relation

$$\text{Wave length in centimeters} = \frac{670d}{\sqrt{E_g}} \tag{180}$$

where d is the diameter of the plate in centimeters and E_g is the grid voltage. It is thus apparent that the higher the frequency to be generated, the greater must be the grid potential and the smaller the dimensions of the tube.

The efficiency of an electron oscillator is low because of the inherent nature of the generating process, with values of 2 or 3 per cent being typical under favorable conditions. The output power obtainable is also low because of the low efficiency and because, with the small dimensions required to generate high frequencies, the allowable power loss at the grid (which must absorb most of the loss) is relatively small. For best output and efficiency it is necessary to adjust carefully the grid voltage, negative plate voltage, and cathode temperature in relationship to the resonant frequency of the external circuit.

Electron oscillations can be produced under a variety of circumstances. The most typical arrangement employs a tube having cylindrical grid and plate electrodes and operates with a slightly negative plate, as explained above, and at a cathode temperature low enough so that very little space charge is formed adjacent to the cathode. It is also possible to obtain electron oscillations with triodes having plane electrodes,[2] with triple-grid tubes,[3] and with plate potentials that are slightly positive.[4] Other modifications are likewise possible, and anyone interested in this subject is referred to the extensive literature available.

[1] This last tendency can be controlled by designing the grid so that it will on the average capture the oscillating electron by the time the phase has shifted sufficiently to cause absorption of energy.

[2] B. J. Thompson and P. D. Zottu, An Electron Oscillator with Plane Electrodes, *Proc. I.R.E.*, vol. 22, p. 1374, December, 1934.

[3] Ferdinand Hamburger, Jr., Electron Oscillations with a Triple-grid Tube, *Proc. I.R.E.*, vol. 22, p. 79, January, 1934.

[4] L. F. Dytrt, Barkhausen-Kurz Oscillator Operation with Positive Plate Potentials, *Proc. I.R.E.*, vol. 23, p. 241, March, 1935.

The resonant circuits used at the extremely high frequencies that can be generated by means of electron oscillations ordinarily consist of resonant lines (Lecher wires) arranged in the form of parallel bars having a length such as to give resonance at the desired frequency.

Electron Oscillators Employing Magnetron Tubes.—Electron oscillations can also be produced by employing a split-anode magnetron tube in the circuit of Fig. 212*a*.[1] When the magnetic field in such a tube is equal to or slightly greater than the value required to prevent current from reaching the plate electrodes, the electrons tend to follow curved paths as illustrated by the dotted paths in Figs. 212*b* and 212*c*, thereby giving rise to electron oscillations having a frequency corresponding to the time required by the electron in its flight.

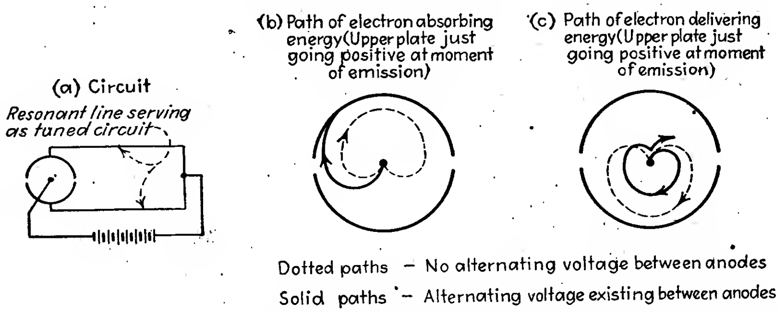


FIG. 212.—Circuit of electron oscillator of split-anode magnetron type, together with electron paths for typical conditions. The magnetic field is perpendicular to the plane of the paper.

The electrons participating in this oscillation tend to synchronize together in much the same manner as with the Barkhausen oscillator. The mechanism involved can be understood by assuming that the magnetic field is just greater than the cutoff value and then postulating a small alternating voltage applied between the two anodes and having a frequency approximating the frequency of the electron oscillations. Consider first an electron emitted as shown in Fig. 212*b*. If the superimposed alternating voltage were zero, this electron would follow the dotted path and would almost, but not quite, reach the anode. If, however, there is a superimposed alternating voltage which is passing through zero and just becoming positive at the instant this electron is emitted, the greater attraction of the upper anode will cause the electron to follow a path such as indicated by the solid line and so reach the anode. In traveling from the cathode to the anode this electron has absorbed a certain amount of energy from the superimposed alternating voltage, and so tends to damp out this oscillation. Now consider an electron emitted

[1] Electron oscillations can also be obtained with a single cylindrical anode, but the efficiency is less than with the split-anode arrangement. See H. Yagi, Beam Transmission of Ultra-short Waves, *Proc. I.R.E.*, vol. 16, p. 715, June, 1928.

at the same instant as the first electron, but from the side of the cathode indicated in Fig. 212*c*. If there were no superimposed alternating voltage, this would follow the dotted path, but because of the superimposed alternating voltage the lower anode is less positive than it would otherwise be. This second electron is attracted less strongly, and hence to follow a path such as indicated by the solid line. As a consequence, this electron does not reach the anode, but circles around and comes to rest near the cathode, after which it starts another cycle of oscillation. This electron delivers energy to the superimposed voltage, and in its complete cycle of oscillation delivers twice as much energy as is absorbed by the first electron. Furthermore, this useful electron is still available for more cycles of oscillation, whereas the first electron is removed at the end of the first half cycle. Electrons leaving the cathode at other times or positions from those considered above give intermediate results. Summarizing, it is seen that the electrons which absorb energy from the superimposed voltage strike the anodes relatively soon and so are removed, whereas those which give up energy to the anode circuit remain for a considerable length of time, with the result that there is a net energy delivered to the anode circuit that tends to sustain the superimposed voltage.

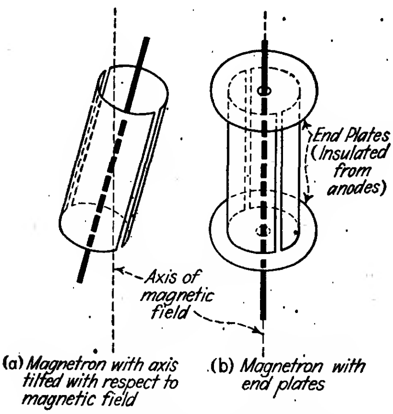


FIG. 213.—Means of causing electrons in magnetron oscillator to spiral out of the interelectrode space.

The behavior of the split-anode magnetron is complicated by the fact that the electron does not travel exactly 360° in one cycle, so that, after a number of cycles have been traversed, the electron that started in such a way as to deliver energy to the superimposed voltage ultimately shifts its phase to the point where it begins to absorb energy. It is hence necessary to remove the electrons from the tube after they have completed a number of oscillations. This can be accomplished by tilting the axis of the tube slightly with respect to the superimposed magnetic field, as illustrated in Fig. 213*a*. Such a tilt gives the electrons a component of velocity parallel to the axis of the tube, causing them to spiral and so ultimately to pass beyond the ends of the anodes. Experiment indicates that the optimum tilt angle is of the order of 3 to 6° in typical tubes.[1] An alternative method of accomplishing the same result

[1] See G. R. Kilgore, Magnetostatic Oscillators for Generation of Ultra-short Waves, *Proc. I.R.E.*, vol. 20, p. 1741, November, 1932.

is to provide the tube with end plates as illustrated in Fig. 213b. By making these end electrodes positive with respect to the cathode, they will attract the electrons, which will follow a spiral path and ultimately be collected by the end plates.[1]

The efficiency of electron oscillators of the split-anode magnetron type is relatively low, values of 5 to 10 per cent being typical under favorable conditions. The power output obtainable is likewise low, particularly at high frequencies, because of the poor efficiency and the fact that the small physical dimensions required to generate energy at high frequencies limit the permissible power dissipation. The best efficiency and power output are obtained when the strength of the magnetic field, the anode voltage, the filament voltage, and the tilt angle (or end-plate potential) are properly adjusted with respect to the resonant frequency of the external circuit.

The wave length λ of a magnetron oscillator is determined primarily by the strength of the magnetic field at which the plate current begins to cut-off according to the following empirical relation[2]

$$\lambda = \frac{13{,}000}{H} \tag{181}$$

where λ is in centimeters and H is in lines per square centimeter (gauss). The critical magnetic field H appearing in Eq. (181) is determined by the dimensions of the tube and the anode voltage, according to the relation

$$H \text{ in lines per square centimeter} = \frac{6.72}{r}\sqrt{E} \tag{182}$$

where r is the anode radius in centimeters and E is the anode voltage. Comparison of Eqs. (181) and (182) shows that small-size and high anode voltages require large H, and hence give high frequencies. The actual frequency is influenced somewhat by the tuning of the external circuit, by space charges developed within the tube, by the tilt angle, etc. As compared with the Barkhausen oscillator, the split anode is capable of operating satisfactorily at higher frequencies because of its simpler structure and greater ability to dissipate energy.[3]

[1] For further discussion of this type of magnetron, see I. Wolff, E. G. Linder, and R. A. Braden, Transmission and Reception of Centimeter Waves, *Proc. I.R.E.*, vol. 23, p. 11, January, 1935; E. G. Linder, Description and Characteristics of the End-plate Magnetron, *Proc. I.R.E.*, vol. 24, p. 633, April, 1936.

[2] For a discussion of the theoretical justification of this equation, see E. C. S. Megaw, Note on the Theory of the Magnetron Oscillator, *Proc. I.R.E.*, vol. 21, p. 1749, December, 1933; J. Barton Hoag, A Note on the Theory of the Magnetron Oscillator, *Proc. I.R.E.*, vol. 21, p. 1132, August, 1933.

[3] The highest frequency reported at this writing is approximately 30,000 mc (1 cm wave length), generated by a split-anode magnetron. See C. E. Cleeton and

Problems

1. In the Hartley and Colpitts oscillator circuits of Fig. 187, follow through the detailed phase relations and show that conditions necessary for oscillation can be realized.

2. Redesign the oscillator in the example of Sec. 66 for a Colpitts circuit.

3. In a properly adjusted oscillator, removing the load resistance normally coupled into the tank circuit will reduce the d-c plate current very greatly, while causing the d-c grid current and the a-c current in the tank circuit to change only slightly. Explain.

4. If an oscillator is suspected of producing intermittent oscillations, what means might be employed to see if this is the case?

5. In an oscillator it is found that the frequency stability is improved by increasing the ratio of reactive energy circulating in the tank circuit to the power output of the oscillator. It is also commonly stated that the use of a tank circuit with a low L/C ratio improves the frequency stability. Correlate these with the discussion on frequency stability given in Sec. 67.

6. In an electron-coupled oscillator, explain why the frequency ceases to be independent of the load impedance in the plate circuit when this impedance is great enough to make the minimum plate voltage very low.

7. Explain how parasitic oscillations can exist in a triode amplifier employing the Rice system of neutralization (see Fig. 133*b*).

8. In a dynatron or other negative resistance oscillator, demonstrate mathematically that oscillations will result when the absolute magnitude of the negative resistance of the tube is equal to or less than the parallel impedance of the tuned circuit.

9. In a three-stage resistance-coupled amplifier, multivibrator oscillations can occur when there is a resistance common to the three plate circuits, whereas oscillations will not normally occur if there are only two stages. Explain how this comes about.

10. Calculate and plot the impedance of the electrical network equivalent to crystal No. III of Table X as a function of frequency, in the vicinity of resonance.

11. *a.* By means of the equivalent electrical circuit of the quartz crystal, demonstrate that when a crystal is used in the oscillator circuit of Fig. 206 the frequency of oscillation can be varied slightly by means of a variable condenser in shunt with the crystal.

b. If the crystal in (*a*) is No. III of Table X, calculate the change in resonant frequency produced by a shunting capacity of 10 $\mu\mu$f.

12. In a crystal oscillator using the circuit of Fig. 206, discuss the factors that control the amplitude of the oscillations produced.

13. Describe the mechanism by which the frequency of an electron oscillator is influenced somewhat by the external tuned circuit, particularly when this circuit is resonant at a frequency not greatly different from the frequency of the electron oscillations in the absence of a tuned circuit.

14. Explain qualitatively the physical reason why the wave length of a Barkhausen oscillation varies with d and E_g as given by Eq. (180).

15. In an electron oscillator of the magnetron type explain why the axial length of the tube in the direction of the filament has no effect on the frequency of oscillation even though a longer tube has more interelectrode capacity.

N. H. Williams, A Magnetostatic Oscillator for the Generation of 1 to 3 Cm Waves, *Phys. Rev.*, vol. 43, p. 421, Sept. 1, 1933.

CHAPTER IX

MODULATION

72. Waves with Amplitude Modulation.—In all the commonly used systems of radio communication the information is transmitted by varying the amplitude of the radiated waves, as explained in Sec. 3 and illustrated in Figs. 4 and 5. Communication carried on in this way is said to take place by means of *amplitude modulation,* and the equation of the envelope of the modulated wave represents the equation of the intelligence actually transmitted. If the envelope variations of the modulated wave exactly reproduce the original signal,[1] *i.e.*, sound pressure, light intensity, etc., this signal is transmitted without distortion; otherwise distortion is introduced and the intelligence contained in the modulation envelope is not exactly the same as that represented by the original signal.

When the equation of the wave envelope does not contain the different frequency components of the original signal in their correct relative magnitudes, the modulation possesses frequency distortion. If on the other hand the equation of the modulation envelope includes frequency components that were not present in the original signal, the modulation possesses amplitude or non-linear distortion. Finally, if the phase relations of the different frequency components of the wave envelope are not the same as in the original signal, phase distortion results. These three types of distortion thus have the same significance when applied to modulation as they do in the case of amplification.

The extent of the amplitude variations in a modulated wave is expressed in terms of the degree of modulation. For a sinusoidal variation, as illustrated by Fig. 5:

$$\text{Degree of modulation} = m = \frac{\text{average envelope amplitude} - \text{minimum envelope amplitude}}{\text{average envelope amplitude}} \tag{183}$$

When the degree of modulation is 1.0, the amplitude variations carry the envelope amplitude to zero during the troughs of the modulation cycle and the modulation is complete, *i.e.*, the envelope is varied through the maximum range that is possible without amplitude distortion.

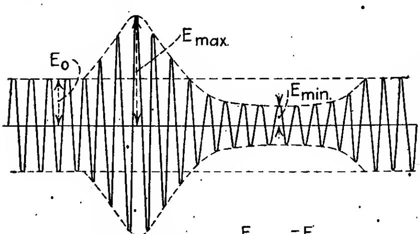
[1] The term "signal" when used in connection with modulation refers to the original intelligence. Thus signal frequency represents the frequency which is modulated upon the radio wave.

When the envelope variation is not sinusoidal, it is necessary to define the degree of modulation separately for the peaks and troughs of the envelope, according to the equations given in Fig. 214.

Analysis of Modulated Wave.—It was shown in Sec. 5 that a modulated wave consists of a carrier wave and a number of side-band components, the exact nature of which depends upon the equation of the wave envelope. This envelope equation can always be written in the form of an average value plus one or more alternating-current terms that take into account the envelope variations, *i.e.*,

$$\text{Equation of wave envelope} = E_0 + E_1 \sin(2\pi f_1 t + \phi_1) + E_2 \sin(2\pi f_2 t + \phi_2) + E_3 \sin(2\pi f_3 t + \phi_3) + \cdots \quad (184)$$

where E_0 is the average amplitude of the envelope (which is also the amplitude when the wave is unmodulated), and the remaining terms represent the components of the envelope variation having frequencies f_1, f_2, f_3, etc., crest amplitudes E_1, E_2, E_3, etc., and phases ϕ_1, ϕ_2, and ϕ_3. The term E_0 in Eq. (184) represents the average amplitude of the modulation envelope and is the crest amplitude of the carrier wave, while each of the sinusoidal components in the wave envelope gives rise to a pair of side-band frequencies. The two side-band components arising from each sinusoidal component of the wave envelope have frequencies that are respectively greater and less than the carrier frequency by the corresponding envelope frequency, and each has an amplitude one-half of the corresponding envelope component. Thus a modulated wave having an envelope given by the equation



$$\text{Positive peak modulation} = \frac{E_{max.} - E_0}{E_0} \times 100$$

$$\text{Negative peak (or trough) modulation} = \frac{E_0 - E_{min.}}{E_0} \times 100$$

FIG. 214.—Unsymmetrically modulated wave, and equations giving peak and trough modulation.

$$\text{Envelope amplitude} = 100 + 50 \sin 2\pi f_1 + 20 \sin 2\pi f_2$$

consists of a carrier wave having a crest value of 100 volts, a side band having a frequency that is f_1 cycles greater than the carrier frequency and a crest amplitude of 25 volts, a companion side band of the same amplitude but of frequency f_1 cycles less than the carrier frequency, and a second pair of side-band components each of 10 volts amplitude and having frequencies f_2 cycles more and f_2 cycles less than the carrier frequency.

The carrier wave is the same, irrespective of the presence or absence of modulation. Varying the amplitude of the radio wave merely has the effect of generating the side bands, which are the part of the modulated

wave that conveys the intelligence being transmitted. The carrier wave carries no information because it is not affected by the modulation.

The energy contained in a modulated wave is the sum of the energies of the separate frequency components and is therefore increased during modulation because of the energy contained in the side bands. When the carrier wave is completely modulated by a sinusoidal variation of the envelope amplitude, *i.e.*, when in Eq. (184) $E_1 = E_0$, and E_2, E_3, etc., are all zero, there are two side-band components, each having an amplitude half that of the carrier and hence each containing one-fourth as much power as does the carrier. The two side bands together thus make the power of the completely modulated wave ($m = 1.0$) 50 per cent greater than the carrier power, and one-third of the total energy of the wave is in the side bands, while two-thirds is in the carrier. With degrees of modulation other than 1.0 the side-band power will be proportional to m^2, so that the fraction of the total wave energy that is contained in the intelligence-bearing side bands rapidly decreases as the degree of modulation is reduced. For this reason the highest possible degree of modulation should always be used.

The frequencies contained in the modulation envelope depend upon the character of the amplitude variations that are impressed on the wave, and in general are higher the more rapidly the amplitude of the wave envelope is varied. In particular when the amplitude changes with extreme rapidity, as would be the case in a code transmitter if the waves could be turned on and off instantly, the envelope equation contains very high frequency components and the side-band frequencies extend over a wide frequency band.

In the actual transmission of intelligence by means of modulated waves it is generally unnecessary to modulate upon the carrier wave all the frequencies contained in the equation of the original signal in order to make the intelligence understandable at the point of reception. Furthermore the modulation by a wider frequency band than is absolutely necessary to carry the desired information is to be avoided because unessential frequencies in the modulation envelope mean high-frequency side-band components, and this utilizes frequencies in the transmitting medium that could otherwise be employed for additional communication.

Side Bands Required in Telegraph, Telephone, and Picture Transmission.—In the transmission of telegraph signals by the Continental Morse Code it is possible to operate telegraph relays provided each side band has a width of 0.131 cycle per letter transmitted per minute.[1]

[1] This figure represents the minimum side band that can be used, and unless refinements are employed a somewhat wider band is desirable. With the five-element two-valued code employed in printing telegraph systems, the side band need be only about three-fourths as wide as with the Continental Morse Code. By employing a

Thus transmission at the rate of 100 letters per minute can be carried on using side bands which extend only 13.1 cycles on each side of the carrier frequency.

In the transmission of speech and music of good quality, side-band components extending at least 5000 cycles on each side of the carrier frequency must be employed. Such a band provides for the transmission of audio-frequency sounds having pitches up to 5000 cycles, and, while the human voice and music contain frequencies up to approximately 15,000 cycles, these higher pitch sounds are not absolutely essential for reasonably satisfactory results. Understandable speech requires the reproduction of all frequencies from about 250 to 2700 cycles, or side-band frequencies ranging from 250 to 2700 cycles above and below the carrier frequency.

The side band required in picture transmission and television is discussed in Sec. 145.

Methods of Amplitude Modulation.—While the methods by which amplitude modulation of a carrier wave can be obtained appear to be almost without number, nearly all of them come under one of the following four classifications:

1. Modulated oscillators.
2. Modulated amplifiers.
3. Modulation by means of non-linear circuit elements.
4. Modulation by means of variable circuit elements.

Modulated oscillators include those arrangements in which the amplitude of the generated oscillations is controlled by the intelligence that is to be transmitted. In the modulated-amplifier type of modulator the unmodulated carrier wave is applied to the amplifier grid, and the amplification is then varied in accordance with the intelligence to be transmitted. Modulators which employ non-linear circuit elements make use of the fact that the current which flows through a non-linear circuit is not proportional to the applied voltage, so that, when the radio-frequency carrier voltage and low-frequency signal voltage are superimposed and applied to a non-linear circuit, the amplitude of the radio-frequency current that flows is modulated by the signal. In modulators employing variable circuit elements, the impedance which is offered to the radio-frequency carrier voltage is varied in accordance with the intelligence that is to be transmitted, with the result that the radio-frequency current is modulated as desired. The following paragraphs describe the methods of amplitude modulation which are most widely used in telephony and

synchronous vibrating relay to restore the shape of the received signals transmitted by the printing telegraph code, it is possible to cut the frequency band in half. See F. E. Terman, Some Possibilities of Intelligence Transmission when Using a Limited Band of Frequencies, *Proc. I.R.E.*, vol. 18, p. 167, January, 1930.

in picture transmission. The means used to turn code transmitters on and off in accordance with the telegraph characters are considered in Sec. 103.

73. Plate-modulated Class C Amplifiers.—A plate-modulated Class C amplifier is an ordinary Class C amplifier in which there is superimposed on the direct-current plate potential an alternating potential that varies in accordance with the intelligence to be transmitted.

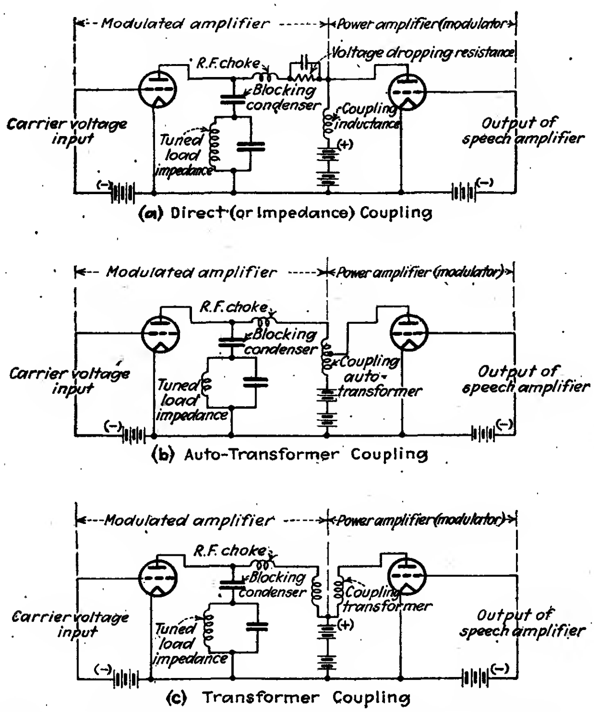


FIG. 215.—Circuit diagrams for plate-modulated Class C amplifier.

The basic circuit of a plate-modulated Class C amplifier is shown in Fig. 215*a*, and is essentially an ordinary triode Class C amplifier with the signal (or modulator) voltage superimposed upon the direct-current plate-supply potential. This modulating voltage is normally obtained from a power amplifier, usually termed the modulator, which can be a Class A, Class AB, or Class B power amplifier.

The most common coupling arrangements employed are illustrated in Fig. 215, which shows schematic circuits which have been simplified by omitting the neutralization and the coupling from tank circuit to load. It will be noted that in each case the plate circuit of the Class C amplifier acts as the load impedance to the modulator. The circuits of (*a*) and

(b) require the use of a single-ended power amplifier, while the arrangement at (c) permits the use of a push-pull modulator and also makes it possible to reduce the direct-current magnetization of the core by polarizing the windings so the direct currents to the Class C and the modulator tube oppose each other in magnetizing the core.

The action of a plate-modulated Class C amplifier is given by the following explanation: The radio-frequency amplifier tube is adjusted so that with no modulating voltage present it operates as an ordinary Class C amplifier in which the bias is at least twice the cut-off value for the plate-supply potential. When the modulating voltage is superim-

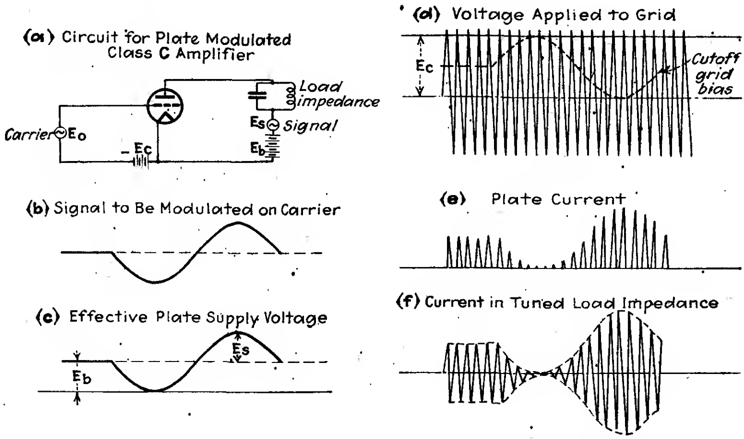


FIG. 216.—Schematic circuit of plate-modulated Class C amplifier, together with oscillograms showing details of the modulator operation.

posed upon the plate-supply voltage, the plate then alternately becomes more and less positive. As the plate becomes more positive, the minimum plate potential E_{min} tends to increase, causing the peak amplitude of the pulse of plate current to become larger, while at the same time the angle of flow increases because of the greater plate voltage. The output is accordingly increased. Similarly when the modulating voltage makes the plate potential less than the plate-supply voltage, the minimum plate voltage E_{min} tends to become less, reducing the peak amplitude of the plate current pulses, and this, together with the smaller angle of flow, reduces the power output. When properly adjusted, the output voltage can be made substantially proportional to the effective plate voltage, and, since this effective plate voltage consists of the battery voltage plus the modulating voltage, the envelope of the output accurately reproduces the signal to give distortionless modulation.

The detailed mechanism of the plate-modulated Class C amplifier is given in Fig. 216, which shows qualitatively the relations existing for

100 per cent modulation. Here the crest value of the signal voltage introduced into the plate circuit is equal to the plate-supply voltage, so that the effective plate-supply voltage varies from zero to twice the battery voltage during the modulation cycle as shown at (c). This causes the plate current pulses to vary in amplitude and angle of flow in the manner shown at (e), resulting in an output current or voltage as illustrated at (f) having an envelope that varies in accordance with the signal.

Power Relations.—When a plate-modulated Class C amplifier is properly adjusted, the plate current and output voltage are nearly exactly proportional to the total effective plate voltage. To the extent that this is true the Class C amplifier offers to the modulator tube a load resistance that is equal to the ratio E_b/I_b, where E_b is the battery voltage and I_b is the d-c plate current in the absence of modulation. The degree of modulation is determined by the alternating voltage developed by the modulator across this load impedance. For 100 per cent modulation the crest alternating signal voltage must equal the plate-supply voltage, and this makes the modulator power $E_bI_b/2$, or exactly one-half of the direct-current power which the plate battery is called upon to supply. For lesser degrees of modulation the modulator power will be proportional to the square of the degree of modulation. *The power relations that exist in a plate-modulated Class C amplifier can hence be summarized by stating that the power required to generate the carrier wave is supplied from the direct-current plate-supply voltage, while the power required to generate the side-band components of the modulated wave must be supplied from the output of the modulator.*

The power that the modulator must deliver becomes large in the case of high-power Class C amplifiers. Thus if a 10-kw carrier is to be modulated 100 per cent and the plate efficiency is 66.7 per cent, the direct-current plate power required is 15 kw and the modulator power output is half of this or 7.5 kw. If the degree of modulation is less than 100 per cent, the demand made upon the modulator is correspondingly less, but so is the side-band energy generated.

The efficiency of a Class C modulated amplifier is the same as the plate efficiency of a simple Class C amplifier. The efficiency is normally at least 60 per cent, and some cases may range as high as 70 to 80 per cent.

Design and Adjustment.—In the modulated Class C amplifier the most important characteristic is the linearity of the modulation, and the entire design and adjustment must be worked out with this in mind. The best means of insuring linear modulation is to adjust the grid excitation and load impedance so that even at the crest of the modulation cycle (when the effective plate voltage is the battery voltage plus the modulated voltage) the tube operates as a Class C amplifier in which the minimum plate voltage is small compared with the total effective plate

voltage. The crest alternating voltage developed across the load is then always almost equal to the effective plate voltage, insuring an output that follows the modulation very closely. It is also highly desirable that the adjustment at the crest of the modulation cycle be such that the maximum grid potential reached during the cycle be considerably less than the minimum plate voltage, so that as the minimum plate voltage becomes smaller at other parts of the modulation cycle it will still exceed the maximum grid potential except at the extreme trough of the modulation cycle. If the exciting voltage is sufficient to make the maximum grid potential larger than this, the grid current will be large during those parts of the modulation cycle when the minimum plate voltage becomes less than the maximum grid potential, while being small during the remainder of the cycle. This causes the modulated amplifier to offer a highly variable load to its exciter.

In order to obtain a reasonable power output with a relatively low minimum plate voltage and a very low maximum grid potential, it is necessary to employ tubes with a low amplification factor. Even then the amplitude of the plate-current pulses and hence the power output tend to be small. When tubes having a relatively high amplification factor are plate modulated, it is usually found necessary to operate with a relatively high maximum grid potential in order to obtain the required output at the peaks of the modulation cycle, and, if the minimum plate voltage is made low to obtain linearity, this makes the maximum grid potential exceed the minimum plate voltage during a considerable part of the modulation cycle and so causes a variable load on the exciter.[1] The effect of the minimum plate potential in relation to the maximum grid voltage upon the linearity of modulation is shown in Fig. 217, where it is seen that, while a relatively low load impedance makes for higher output power and more uniform driving power over the cycle, the linearity of modulation is less than with a higher load impedance.

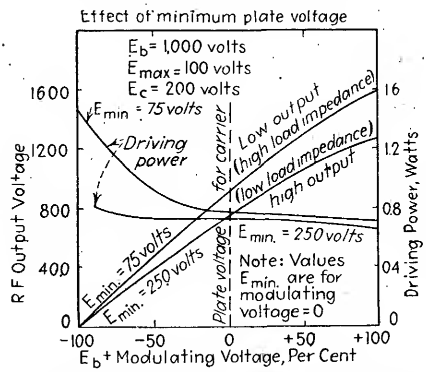


FIG. 217.—Effect of various operating conditions on the linearity and driving power of a plate-modulated Class C amplifier.

[1] This difficulty can be minimized by several expedients such as modulating the exciter slightly, applying a small amount of grid modulation to the plate-modulated amplifier, or using grid-leak bias. In this connection see I. E. Mouromtseff and H. N. Kozanowski, Analysis of the Operation of Vacuum Tubes as Class C Amplifiers, *Proc. I.R.E.*, vol. 23, p. 752, July, 1935.

In designing a modulated Class C amplifier it is customary to start with a grid bias that is at least twice, and more commonly three to four times, the cut-off value calculated on the basis of the plate-supply voltage. An excitation voltage and load impedance are then selected such that with the effective plate voltage existing at the crest of the modulation cycle (which for 100 per cent modulation is twice the plate-supply voltage) the maximum grid potential and minimum plate voltage will have values favorable for linear modulation while still giving reasonable power output. The proper conditions can be readily calculated with an accuracy sufficient for ordinary purposes by the approximate method discussed in Sec. 61. The effective plate voltage at the crest of the modulation cycle divided by the calculated d-c plate current at the same instant gives the approximate load that the modulated amplifier offers to the modulator tube, and from this the coupling arrangement can be designed. In the case of the direct-coupled circuit of Fig. 215*a* it is necessary to insert a voltage-dropping resistance, thoroughly by-passed to both audio and radio frequencies, in series with the plate of the Class C amplifier as shown if 100 per cent modulation is to be obtained. This is because the alternating voltage developed by the modulator tube across its load impedance is always less than the modulator plate voltage, and makes it necessary to lower the Class C amplifier direct-current plate voltage by a corresponding amount, usually 20 to 40 per cent.

In selecting operating conditions for a plate-modulated amplifier it is necessary to remember that with complete modulation the power input and plate losses are approximately 50 per cent greater than with no modulation, while the peak plate voltage is twice as great. Hence, when a tube is to be plate modulated, it is necessary to operate at a lower plate voltage and to adjust for less carrier power than could be developed by the same tube if operated as an ordinary Class C amplifier without modulation. In the case of tungsten filament tubes it is also necessary to limit the peak plate currents to about two-thirds of the value permissible with Class C operation to avoid incipient saturation effects.

After adjusting to realize the operating conditions tentatively selected on the basis of the above considerations, modulating voltage is applied and the linearity determined either by employing a modulation meter to read the percentage modulation for the positive and negative peaks or by using a cathode-ray tube. The modulation meter is the more accurate, but, though desirable where high quality is important, is expensive and often not available. The cathode-ray tube is not so accurate as the modulation meter but because of its general availability is used for the ordinary run of adjustments. The simplest procedure when employing a cathode-ray tube is to apply the signal or modulating voltage to the horizontal deflecting plates and the modulated output

voltage to the vertical deflectors. If the modulation is distortionless, the resulting pattern is a trapezoid with straight sides as indicated in Fig. 218, with the degree of modulation determined by the ratio of the vertical sides as indicated in the figure. Overmodulation, *i.e.*, modulation to the point where the output is reduced to zero for an appreciable part of the cycle, is indicated by the pattern shown at Fig. 218*b*. Amplitude distortion causes the sides of the trapezoid to become curved as in Fig. 218*c*, with the nature of the curvature usually indicative of the source of the distortion. If a shaded ellipse is indicated as in Fig. 218*d*, a phase shift is present.

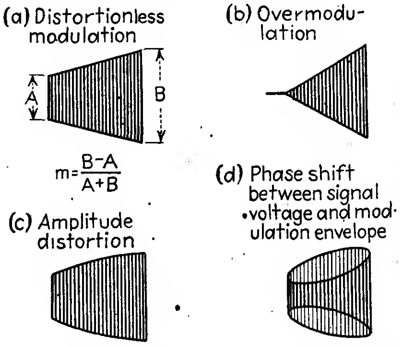


FIG. 218.—Patterns obtained under various conditions from a cathode-ray tube in which the modulated wave is applied to the vertical delecting plate and the signal voltage to the horizontal deflectors.

In the absence of a modulation meter or cathode-ray tube, an idea as to the linearity of the modulation may be obtained by taking advantage of the fact that when the modulation is distortionless the d-c plate current is substantially independent of modulation, whereas distortion is usually accompanied by a change in the d-c current when modulation is applied. The percentage of modulation can be estimated for sinusoidal modulation by taking advantage of the fact that the effective value of a sinusoidally modulated wave is proportional to $\sqrt{1 + \frac{m^2}{2}}$, and so with complete modulation is 1.225 times the unmodulated value.

If the linearity of modulation, power output, or grid driving power for the first trial adjustment are not satisfactory, the exciting voltage, load impedance, or grid bias can be varied in an effort to improve the performance. The practical problem of determining the optimum adjustment is complicated by the variable load that the modulated amplifier offers to its exciter and by the fact that the power output of a modulated amplifier is commonly used to excite a linear amplifier which also represents a variable load. In some cases it is possible to balance some of these modifying factors against each other to make the performance better than otherwise obtainable. Thus in the arrangements illustrated in Fig. 215*a*, where a single-ended power amplifier must be employed, it is often possible to adjust the modulated amplifier so that it compensates to a large degree for the second-harmonic distortion always present in a single-ended power amplifier.

The tank circuit of a modulated amplifier must have a Q low enough to prevent undue discrimination against the higher side-band frequencies. When a high load impedance is required, it is therefore necessary to make the L/C ratio of the tank circuit high enough to permit realizing the required impedance with the relatively low effective Q necessary for high quality.

Calculation of Performance.—In calculating the linearity of a plate-modulated Class C amplifier it is necessary to determine the alternating voltage developed across the load for known values of grid bias, maximum grid potential, effective plate voltage, and load impedance. To do this

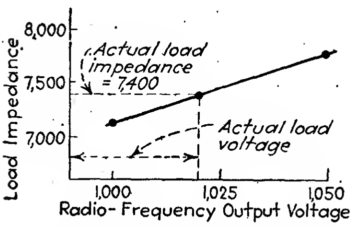


FIG. 219.—Curve showing load impedance required for different assumed values of minimum plate voltage E_{min}, showing how the value of E_{min} corresponding to a given load impedance can be found.

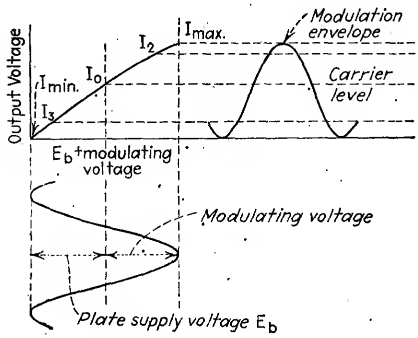


FIG. 219*a*.—Critical points on linearity curve used to calculate distortion in modulation by means of Eq. (151) or (152).

it is necessary to assume reasonable values for the minimum plate potential E_{min} and then to calculate the resulting current pulses and the required load impedance that is consistent with these current pulses and the effective plate voltage. By plotting several such points in a curve showing the required load impedance as a function of the minimum plate voltage E_{min} as in Fig. 219, it is possible to determine the value of minimum plate voltage corresponding to the actual load impedance, and from this to obtain the voltage developed across the load. This calculation can be repeated for various values of effective plate voltage in order to obtain a complete curve giving the output voltage as a function of effective plate voltage. Calculations of the linearity must be carried out by the exact method (see Sec. 61) if high accuracy is required, although the approximate method is suitable for preliminary calculations or for detecting grossly improper operating conditions.

When the linearity curve is known, the amplitude distortion of the modulation envelope can be determined from the output at the critical

points on the modulation cycle indicated in Fig. 219*a* by substitution in Eq. (151) or (152).

Plate Modulation of Screen-grid and Pentode Tubes.[1]—Pure plate modulation of screen-grid and pentode tubes is not desirable because the plate current of such tubes is substantially independent of the plate voltage except when the plate potential is quite low. Entirely satisfactory modulation can be obtained, however, if both plate and screen potentials are modulated simultaneously as in Fig. 220. In this way the total space current varies with the modulation in much the same way as in the ordinary plate-modulated amplifier. The design and adjustment of such a modulation system is similar to that of the ordinary plate-modulated triode except that the screen-grid by-pass condenser must be small enough so that it does not by-pass the modulating voltage.

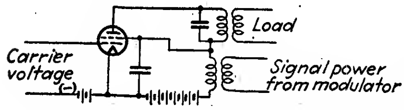


FIG. 220.—Modulated screen-grid amplifier in which the screen and plate voltages are simultaneously modulated.

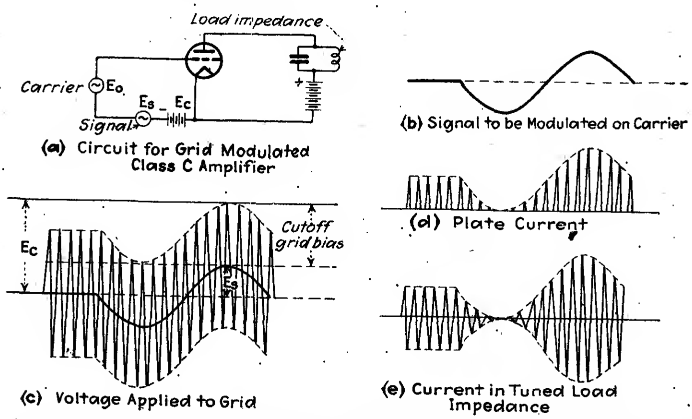


FIG. 221.—Circuit of grid-modulated Class C amplifier together with oscillograms showing details of operation.

74. Grid-modulated Class C Amplifier.—The grid-modulated Class C amplifier is similar to the plate-modulated Class C amplifier except that the modulation is accomplished by varying the grid bias of the amplifier instead of the plate voltage. The circuit for such an amplifier together with the details of operation are illustrated in Fig. 221. The tube is operated at a grid bias greater than the cut-off value, and the amplitude of the carrier wave that is applied to the grid is somewhat less than the bias. A signal voltage is then superimposed on the bias and causes

[1] For further information see H. A. Robinson, An Experimental Study of the Tetrode as a Modulated Radio-frequency Amplifier, *Proc. I.R.E.*, vol. 20, p. 131, January, 1932.

the voltage actually applied to the grid to vary as shown in Fig. 221*c*. The result is the plate-current impulses shown in Fig. 221*d*, which are similar to the corresponding plate-current impulses of the plate-modulated amplifier.

The proper adjustment is such that at the crest of the modulation cycle the tube operates under typical Class C amplifier conditions, with plate-current pulses lasting for not more than 180 deg., a maximum instantaneous grid potential equal to zero or moderately positive, and sufficient load impedance to make the alternating voltage developed across the load by the plate-current pulses only slightly less than the battery voltage. Complete modulation is then obtained by arranging matters so that at the trough of the modulation cycle the maximum instantaneous grid voltage approaches the cut-off grid potential for the plate-supply voltage.

In actual practice it is necessary to compromise between power output and distortionless modulation. In order to obtain a large power output the pulses of plate current at the crest of the modulation cycle must be large. This requires that the grid be driven positive; however, the resulting grid current that flows at the peak of the modulation cycle places a variable load upon the exciting and modulating voltages, thereby distorting the modulation. By designing the sources of exciting and modulating voltages to have good regulation (*i.e.* low internal impedance) and by driving the grid only moderately positive, this distortion can be kept low enough to be permissible for such purposes as police radio, aircraft telephone, etc., where the transmission of intelligence is the primary object. However, if the distortion is to be kept as low as possible, as is essential in broadcast transmitters, operation must be as illustrated in Fig. 221 with the maximum instantaneous grid potential just reaching zero. It is also desirable to use sufficient load impedance to make the minimum plate potential E_{min} small compared with the plate voltage. In order to obtain appreciable power output with the grid just reaching zero potential and the minimum plate voltage quite low, it is necessary to employ a tube with a low amplification factor. Even then the power output obtainable at the crest of the modulation cycle will usually not bring the plate dissipation of the tube up to the rated value, so that, in order to obtain a high degree of linearity, it is necessary to sacrifice in power output.

The first step in adjusting a grid-modulated amplifier is to obtain the proper operating conditions for the crest of the modulation cycle. This is done by temporarily giving the grid a bias equal to or slightly greater than cut-off and by applying sufficient exciting voltage to make the maximum grid potential either zero or moderately positive according to the type of operation desired. The plate load impedance is then

adjusted, keeping in mind that the modulation will be more linear as the load impedance becomes higher, but that this tends to reduce the power output. The grid bias is now increased until the output drops to zero, and the operating grid bias is selected midway between this bias and that corresponding to the crest of the modulation cycle. The crest modulating voltage for 100 per cent modulation is then equal to one-half the difference between the biases for zero output and peak output, and ordinarily will be very close to $E_b/2\mu$.

The linearity of the modulation can be determined experimentally by measuring the output voltage as a function of grid bias. Such a procedure takes into account the effect of the grid current upon the exciting

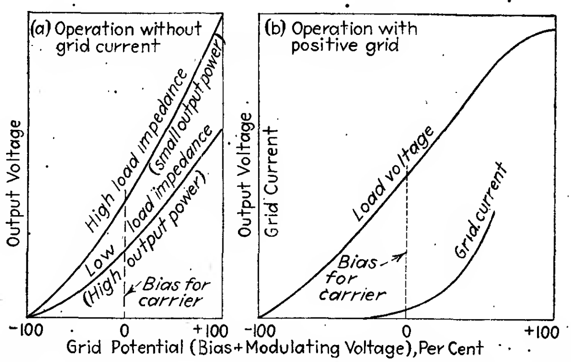


FIG. 222.—Typical linearity curves of grid-modulated amplifier for several conditions.

voltage, but does not allow for the flattening off of the positive peaks of the modulating voltage, which must be taken into account by the use of Eq. (156). Typical linearity curves for several operating conditions are illustrated in Fig. 222. It is apparent that, if the load impedance is high and the grid is not driven positive, the modulation can be made very linear, but that with a low load impedance or with positive grid operation, the resulting increase in output is at the expense of linearity or of variable driving power.[1]

The performance of a grid-modulated Class C amplifier can be calculated by the exact or approximate methods of Sec. 61 in much the same manner as the calculations for the plate-modulated Class C amplifier

[1] It is to be noted by comparing Figs. 152 and 222 that operation with zero grid current and a low load impedance gives a characteristic that is curved in much the same way as is the dynamic characteristic of a power tube. Hence by using a single-ended power tube to supply the modulating voltage it is possible by proper adjustment to balance the curvature of the dynamic characteristic of the power tube against the curvature of the modulation characteristic of the modulated amplifier and to obtain as a result a substantially distortionless over-all characteristic with a load impedance low enough to give appreciable output.

are carried out. Thus, for any particular value of modulating voltage, various values of minimum plate voltage are assumed and the required load impedance is calculated. From these data a curve is plotted showing the relationship between minimum plate voltage and required load impedance, and the conditions corresponding to the actual load impedance are obtained. Such calculations using the approximate method are very helpful in preliminary design to determine the proper load impedance, output power that can be expected, etc.; in the case where no grid current flows, they will also give the linearity curve with satisfactory precision. When the grid is driven positive, the distortion is greatly influenced by the grid current at the peak of the cycle and linearity under such conditions is best determined by experiment.

The carrier power obtainable from a grid-modulated amplifier is approximately one-quarter of the power that the same tube will deliver operating as a simple Class C amplifier. This is because the peak power of the modulated amplifier corresponds to Class C operation, and with a completely modulated wave is four times the carrier power. The plate efficiency during the unmodulated intervals is approximately one-half the efficiency obtained with simple Class C operation. This results from the fact that, if the minimum plate voltage is small at the crest of the modulation cycle, then when there is no modulation the voltage across the load is halved, causing the minimum plate potential and hence the plate loss to become large.

Upon comparing the grid- and plate-modulated amplifiers it will be noted that the grid-modulated amplifier has a relatively low plate efficiency and a low power output in proportion to the size of the modulated tube, but requires very little modulating power. In contrast with this the plate-modulated amplifier gives a much larger output in proportion to the size of the modulated tube, but requires a large modulator power. As a consequence the over-all efficiency, considering both modulator and modulated-tube capacities, is roughly the same. The choice between the two methods of modulation is hence largely one of convenience since both methods of modulation will give completely modulated waves having negligible distortion.

Grid Modulation with Screen-grid and Pentode. Class C Amplifiers.—Grid modulation can be readily applied to screen-grid and pentode Class C amplifiers. The principles involved in the operation design and adjustment are much the same as when triodes are used except that no neutralization is needed.

75. Suppressor-grid Modulation.—This method of modulation makes use of a pentode tube adjusted to operate as a Class C amplifier. Modulation is accomplished by applying to the suppressor grid the modulating signal voltage superimposed upon a suitable negative bias, as illustrated

in Fig. 223. This arrangement takes advantage of the fact that the lowest plate potential at which appreciable plate current starts to flow is proportional to the suppressor-grid potential, as indicated in Fig. 65. If the load impedance in the plate circuit of the tube is high, then the minimum plate voltage will always be only slightly more than the value at which the plate current approaches cut-off. This makes the output voltage follow variations in the suppressor-grid potential very closely, insuring relatively linear modulation.

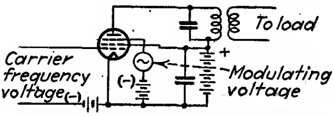


FIG. 223.—Circuit of suppressor-grid modulated amplifier.

In the practical adjustment of suppressor-grid modulated amplifiers the first step is to obtain the proper control-grid bias, screen-grid potential, and excitation to give the required angle of flow and total space current, just as in the case of a Class C amplifier. The load impedance is then adjusted to the highest value at which most of this total space current goes to the plate when the suppressor is slightly positive, but still

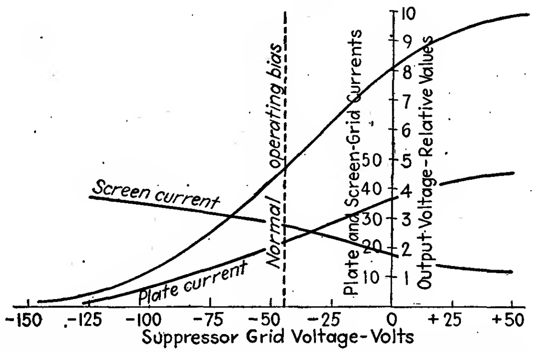


FIG. 224.—Linearity curve of typical suppressor-grid modulated amplifier. Note that it is impossible to obtain complete modulation without introducing appreciable distortion.

not sufficiently positive to draw very much current. This condition corresponds to the crest of the modulation cycle. The suppressor potential required to reduce the output to substantially zero is then determined experimentally, thereby obtaining the bias voltage and the modulating voltage required for complete modulation.

The power output of a suppressor-modulated amplifier at the crest of the modulation cycle corresponds to the output obtainable from the same tube operating as a Class C amplifier, so that for 100 per cent modulation the carrier power is approximately one-fourth of that obtainable from the same tube when used as a simple amplifier. The plate efficiency when adjusted for complete modulation is the same as in the grid-modulated amplifier, *i.e.*, approximately one-half the plate efficiency

for Class C operation, and is low for the same reason. The total efficiency including the screen-grid losses is lower with suppressor modulation than with control-grid modulation because with suppressor-grid modulation the total space current is practically constant throughout the modulation cycle, resulting in a large screen current at the troughs of the modulation cycle and consequent high power loss at the screen grid. The relations involved are illustrated in Fig. 224, which shows the plate current, screen current, and output voltage as a function of suppressor voltage in a typical case.

It will be noted that suppressor-grid modulation is very similar to control-grid modulation, the primary difference being that the exciting and modulating voltages are applied to different grids. The linearity

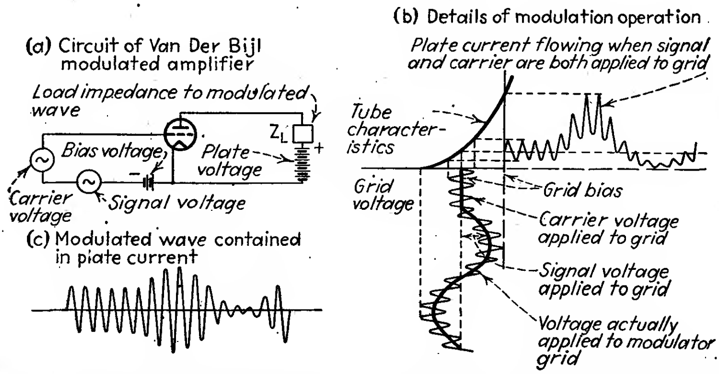


FIG. 225.—Circuit of van der Bijl type of modulated Class A amplifier, together with oscillograms showing details of operation. The curvature of the grid-voltage-plate-current tube characteristic causes the amplification of the carrier to depend upon the grid potential, which in turn varies in accordance with the signal.

obtained with suppressor modulation is about the same as with control-grid modulation when the control grid is driven positive, but is decidedly less than the linearity obtainable with the control-grid modulation operated without grid current.

76. The van der Bijl Type of Modulated Class A Amplifier.—This type of modulator consists of an ordinary Class A power amplifier, to the grid of which are applied a small radio-frequency carrier voltage and a large signal voltage. Because of the curvature of the plate-current–grid-voltage characteristic the amplification of the small carrier voltage depends upon the amplitude of the signal voltage, thus causing the amplified output to be the desired modulated wave. The detailed mechanism by which the modulation is produced is shown in Fig. 225. The degree of modulation depends upon the curvature of the $E_g - I_p$ characteristic and the amplitude of the signal voltage that is applied to the grid. In order to obtain a high degree of modulation the tube must

be operated over a curved portion of the $E_g - I_p$ characteristic, and the signal voltage must be large.

An analysis of the van der Bijl modulator can be readily made by means of the complete equivalent circuit of the vacuum-tube amplifier illustrated in Fig. 151 and discussed in Sec. 55. The side bands of the modulated output are the sum and difference frequency terms produced by the second-order action when the input consists of two sine waves [see Eq. (135b)], while the carrier is the carrier-frequency output produced by first-order action. Analysis based on the equivalent circuit shows that for best results the load impedance in the plate circuit should have negligi-

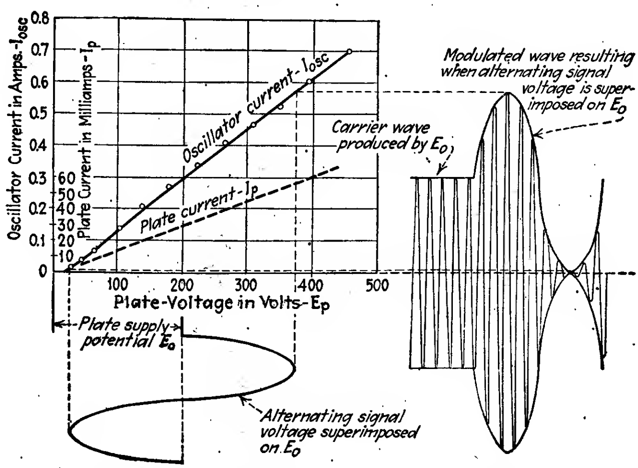


FIG. 226.—Plate and resonant-circuit currents as a function of plate-supply voltage in a typical oscillator having grid leak and condenser, and adjusted for efficient operation. The oscillating current is seen to be almost exactly proportional to the plate voltage, so that the oscillations can be modulated by varying the plate-supply voltage about the average value E_0 as shown in the figure.

ble impedance to the signal frequency and at the same time offer to the modulated wave a resistance equal to one-third the plate resistance at the operating point.

The van der Bijl type of modulated amplifier is relatively easy to adjust and requires negligible signal power. At the same time, its usefulness is limited by the fact that the plate efficiency, *i.e.*, the ratio of power contained in the modulated wave to the direct-current plate power, is very low and by the fact that a high degree of distortionless modulation is hard to obtain. This system of modulation is used extensively in carrier-current telephone communication, where only small amounts of modulated power are required and where a high degree of modulation is not necessary because the carrier component is suppressed by the means described in Sec. 80.

77. Plate-modulated Oscillator.—In this method of modulation advantage is taken of the fact that in an oscillator adjusted to operate at good efficiency the alternating voltage developed across the tuned circuit is almost exactly proportional to the plate potential as explained in Sec. 66 and illustrated in Fig. 226. It is hence possible to modulate the output of such an oscillator by superimposing the signal upon the plate-

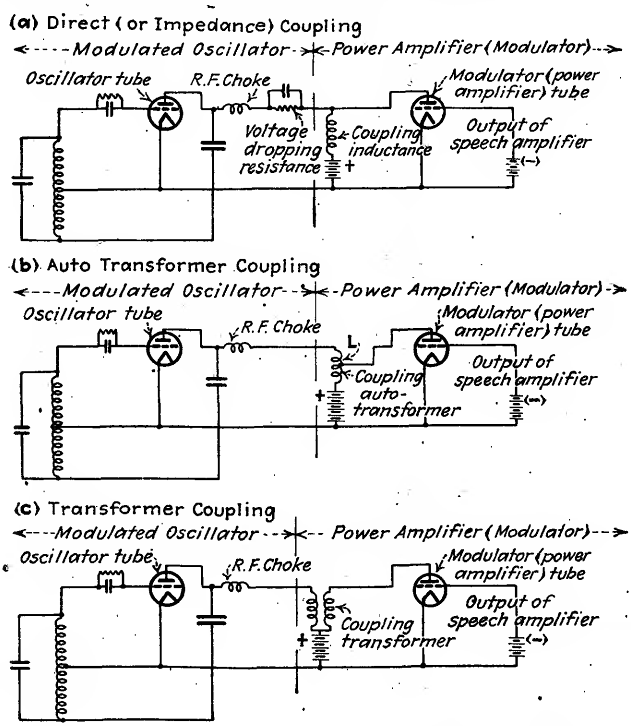


FIG. 227.—Circuits for the plate-modulated oscillator. In each case the oscillator plate circuit represents the load impedance to which the amplifier power output is delivered.

supply potential by coupling a power amplifier (or modulator) to the oscillator in one of the arrangements illustrated in Fig. 227.

The power relations existing in a plate-modulated oscillator are similar to those of the plate-modulated Class C amplifier. The plate battery is hence called upon to supply the power for generating the carrier while the modulator must deliver sufficient power to generate the side-band energy. The plate efficiency is the same as that of an ordinary oscillator, and so ranges commonly from 60 to 80 per cent.

The linearity of a properly adjusted plate-modulated oscillator is practically perfect because under such conditions the minimum plate potential reached during the radio-frequency cycle is small and approximately proportional to the effective plate-supply voltage, making the

voltage developed across the load always proportional to the plate potential. The only special precaution required to obtain good linearity, other than arranging the tube to operate at high plate efficiency and to have an effective tank-circuit Q that is not too high, is to employ a small enough grid condenser so that the grid bias is capable of following the modulation at the higher modulation frequencies. In order to avoid distortion from this source, it is essential that the grid condenser reactance at the highest modulation frequency be at least several times the grid-leak resistance if 100 per cent modulation is to be employed. The chief limitation of this type of modulation arises from the fact that the frequency generated by ordinary oscillators depends somewhat upon the plate-supply voltage. The carrier frequency generated by a plate-modulated oscillator therefore tends to vary with the modulation envelope, introducing frequency modulation (see Sec. 81). As a result the plate-modulated oscillator, which was once the universally used method of modulation, is now employed only under special circumstances.

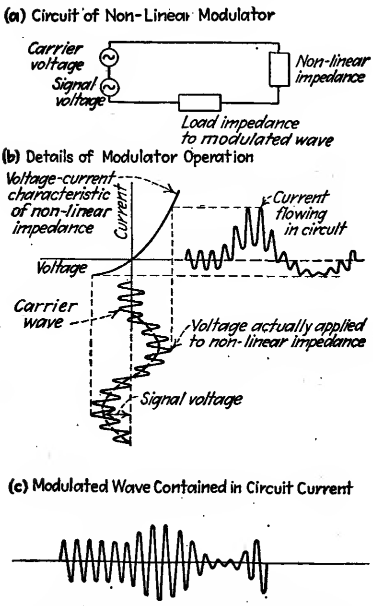


FIG. 228.—Circuit of modulator utilizing a non-linear circuit element, together with details showing mechanism of operation. The signal voltage varies the impedance offered to the carrier voltage and thus causes the current that flows to be a modulated wave.

78. Miscellaneous Methods of Modulation.—While the preceding sections describe the most common types of modulators, many other methods have been devised. The most important and representative of these miscellaneous modulation methods are briefly considered in the following paragraphs.[1]

Modulators Employing a Non-linear Impedance.—Modulators utilizing non-linear impedances operate by superimposing the signal and carrier voltage upon each other and applying the combination to a circuit

[1] For still other methods of modulation see:

Charles A. Culver, An Improved System of Modulation in Radio Telephony, *Proc. I.R.E.*, vol. 11, p. 479, October, 1923; and Series Modulation, *Proc. I.R.E.*, vol. 23, p. 481, May, 1935.

Ernest G. Linder and Irving Wolff, Note on an Ionized Gas Modulator for Short Radio Waves, *Proc. I.R.E.*, vol. 22, p. 791, June, 1934.

For the modulation of laboratory oscillators, see the author's book "Measurements in Radio Engineering," 1st ed., pp. 294–298.

element having an impedance that depends upon the applied voltage. The action that takes place is shown in Fig. 228 and in a broad way can be summarized by stating that the carrier-frequency current that flows depends upon the amplitude of the superimposed signal-frequency voltage, with the result that a modulated wave is produced. The only essential difference between the non-linear impedance and van der Bijl modulators is that the latter involves amplification. Non-linear modulation is often unintentionally produced in vacuum-tube amplifiers by the curvature of the plate-voltage–plate-current characteristic. A practical application of the non-linear circuit-element method of modulation is in the grid-current modulator shown in Fig. 229,[1] which makes use of the non-linear resistance that exists between grid and cathode of a vacuum tube operated at zero grid bias. The signal and carrier voltages are applied to the grid of the tube, and as a result of the non-linear grid-cathode tube resistance a modulated carrier wave flows in the grid circuit. This wave is forced to flow through the grid impedance Z_g, across which there is produced a voltage drop that is amplified in the plate circuit of the tube by ordinary amplifier action.

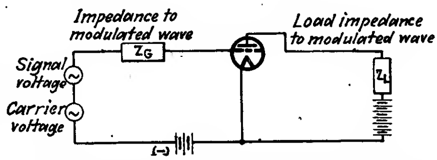


FIG. 229.—Grid-current modulator, in which the non-linear relation between grid current and grid voltage is used to produce a modulated wave of grid current that, in flowing through the series grid impedance, produces a voltage drop which is then amplified by the tube.

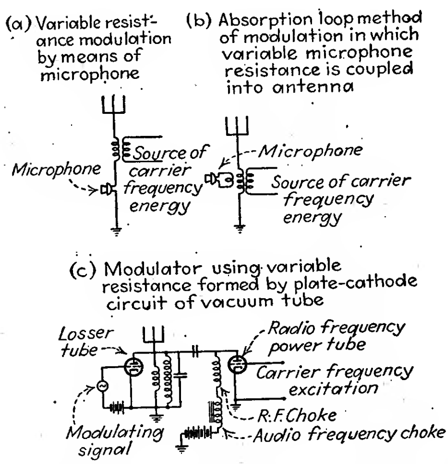


FIG. 230.—Representative modulation methods employing a circuit element having an impedance that depends on the signal amplitude.

Non-linear modulators employing thyrite blocks and copper-oxide rectifier units to supply the non-linear circuit arrangement find use in telephone and similar applications where the amount of modulated power required is quite small and the degree of modulation required is not necessarily high.

Modulation by Means of a Variable Impedance.—A number of methods of modulation which operate by varying an impedance in accordance with

[1] See Eugene Peterson and Clyde R. Keith, Grid Current Modulation, *Bell System Tech. Jour.*, vol. 7, p. 106, January, 1928.

the intelligence to be transmitted are shown in Fig. 230. The circuit shown at Fig. 230*a*, which is probably the first modulation system ever devised, operates by varying the resistance of the antenna circuit in accordance with the resistance of a carbon-type microphone. A modification of this circuit is shown at Fig. 230*b*, in which the microphone is inductively coupled to the antenna system. Both of these arrangements can be used to modulate small amounts of power but at best are rather unsatisfactory.

A recently developed absorption modulator system is shown in Fig. 230*c*. This makes use of a losser tube arranged so that the power absorbed is controlled by the potential applied to the losser grid. An essential feature of the arrangement is the choke in series with the plate of the radio-frequency power tube. This choke must offer a high impedance to the modulating frequency, and tends to maintain the plate current of the power tube constant, thereby improving both the over-all efficiency of the modulating system and the degree and linearity of modulation obtainable.

79. Comparison of Modulation Methods.—An idea as to the advantages, disadvantages, and fields of usefulness of the various systems of modulation can be gained from the following discussion.

The plate-modulated oscillator is simple, has good plate efficiency, and is the most linear of all modulation systems, but is of limited usefulness because of the frequency modulation produced by the plate-voltage variations. Hence this system of modulation is employed only where simplicity is all important, and in laboratory and test oscillators.

The plate-modulated Class C amplifier is the method of modulation most widely used in radio work. The linearity can be made sufficiently high to meet all practical requirements, and the plate efficiency is high. The chief disadvantage is the large amount of modulating power required, which is an especial disadvantage where amplification of the signal frequencies is difficult, as in television.

Control-grid modulation has the advantage of requiring negligible modulating power, but this is counteracted by low plate efficiency, with the result that, when the total installed equipment, including both the modulated amplifier and modulator tube, is considered, there is little difference in economy between control-grid and plate modulation. When the control-grid modulator is operated without drawing grid current, the linearity is of the same character as obtainable with plate modulation, but the output produced by a given tube tends to be low. The output can be increased by driving the control grid slightly positive, but this introduces more distortion than is ordinarily permissible in broadcast transmitters, although the performance is still satisfactory for such services as police radio where the transmission of information rather than entertainment is the objective.

Suppressor-grid modulation corresponds very closely to control-grid modulation in which the control grid is driven positive. The power output, linearity, and plate efficiency are of the same order of magnitude, with suppressor-grid modulation having the advantage of requiring no neutralization and of being simpler to adjust.

The van der Bijl modulated amplifier is widely used in the balanced-modulator circuit of Fig. 231 to produce single side-band energy for carrier telephone communication, but is not used where appreciable power is required because of the low efficiency and the fact that it is difficult to obtain 100 per cent modulation without excessive distortion. Non-linear modulators employing thyrite or copper oxide have displaced the van der Bijl modulator to some extent in recent years.

Absorption systems of modulation are no longer employed, although the recent arrangement of Fig. 230*c* has practical possibilities, particularly as it can readily be applied to existing code transmitters.

80. Carrier-suppression and Single Side-band Systems of Communication.—The carrier component of a modulated wave has frequency, amplitude, and phase that are not affected by the presence or absence of modulation, and so contains no part of the intelligence being transmitted. The carrier wave can therefore be suppressed at the transmitter by some arrangement such as the balanced modulator circuit shown in Fig. 231. Here the carrier voltage is applied to the two modulator tubes in the same phase, as shown in the figure, while the modulator signal voltage is applied in opposite phase to the two grids by means of the center-tapped transformer. The outputs in the plate circuits of the two tubes are combined through a transformer with a center-tapped primary in such a way that voltages applied to the two grids in the same phase cancel each other in the output, while voltages applied to the two grids 180° out of phase are added in the output. The result is that the carrier voltage, which is applied to the two tubes in the same phase, does not appear in the secondary of the output transformer, whereas the side-band components, which are produced in opposite phase, are added to give an output that is a modulated wave from which the carrier component has been removed.

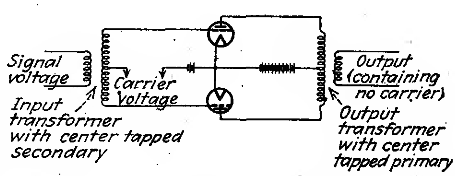


Fig. 231.—Balanced modulator circuit arranged to suppress the carrier wave from the output without altering the side bands.

Since each side band taken alone contains all the information present in a modulated wave, it is possible to carry on communication by transmitting only a single side band and by suppressing the carrier and other side band at the transmitter. The single side band is obtained by first suppressing the carrier by some such arrangement as shown in Fig. 231

and by then passing the resulting side bands through filter circuits that are sufficiently selective to transmit one side band while suppressing the other.

The single side-band system of communication is able to transmit a given signal with a frequency band only half as wide as that required by a modulated wave consisting of two side bands and a carrier, and also saves over two-thirds in power because of the suppression of the carrier.

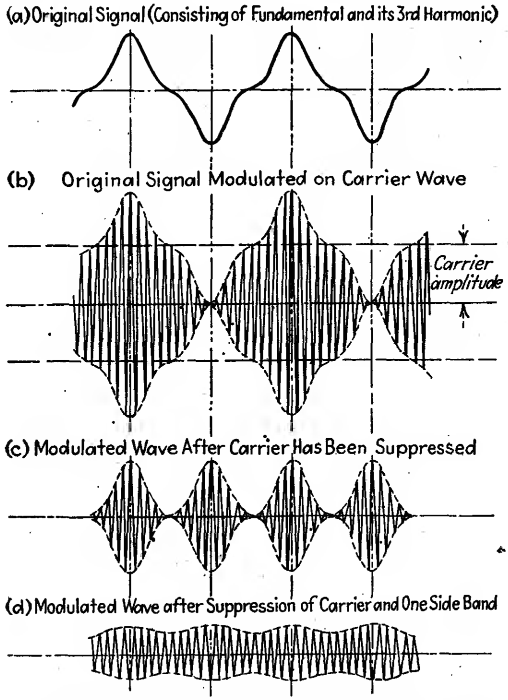


FIG. 232.—Character of waves produced when carrier is suppressed, and when carrier and one side band are both suppressed. When the carrier is removed, the envelope varies at a frequency that is twice that when the carrier is present, while, when only a single side band is present, the envelope varies in accordance with the difference frequencies formed by the components of the original signal, which in the case shown consisted of a fundamental wave and its third harmonic.

Single side-band transmission is extensively used in carrier-current communication over wire lines, but the difficulty of producing large amounts of single side-band power at radio frequencies and the difficulty of receiving the signals have prevented single side-band transmission from being standard practice in radio work.

Signals in which the carrier, or the carrier and one side band, have been suppressed are received by combining them with a locally generated oscillation having a frequency as close as possible to that of the original carrier. In the case of single side-band transmission results are satis-

factory if the local oscillation is within 5 to 10 cycles of the correct frequency. However, when both side bands are transmitted, the local oscillation must have exactly the proper frequency and even the correct phase, which makes double side-band carrier-suppression systems of transmission entirely impractical.[1]

The waves that are sent out by carrier-suppression and single side-band systems of communication differ in appearance from a modulated wave in several respects as is apparent from Fig. 232, which shows the same signal transmitted by amplitude-modulation, carrier-suppression, and single side-band systems. The wave with carrier suppression differs from the modulated wave primarily in having an envelope that varies in

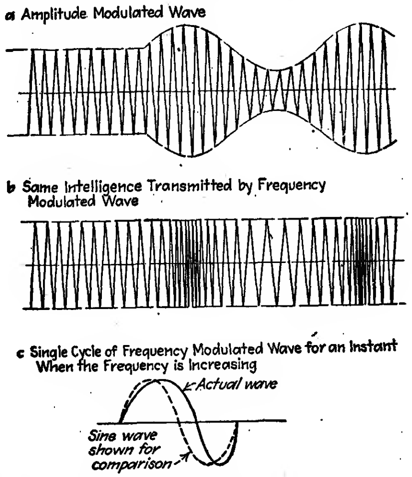


FIG. 233.—Character of waves produced by frequency modulation, together with large-scale reproduction of a single cycle, showing how the wave shapes are not sinusoidal.

amplitude at twice the modulation frequency as a result of action between the two side bands. The wave representing a single side band consists of a number of frequency components, one for each component in the original signal. Each of these components has an amplitude proportional to the amplitude of the corresponding signal component and a frequency differing from that of the carrier by the signal frequency. The result is that in a general way the envelope amplitude of the single side-band signal increases with the degree of modulation and varies in accordance with the difference frequencies formed by the various frequency components of the single side-band interacting with each other.

81. Frequency and Phase Modulation.—Instead of transmitting intelligence by varying the amplitude of the radiated wave, it is also possible to carry on communication by keeping the amplitude of the

[1] See R. V. L. Hartley, Relations of Carrier and Side Bands in Radio Transmission, *Proc. I.R.E.*, vol. 11, p. 34, February, 1923.

wave constant and varying the frequency in accordance with the signal to be transmitted. This is known as frequency modulation and results in a radiated wave having the appearance shown in Fig. 233*b*. The extent of the frequency variation in such a wave is made proportional to the amplitude of the signal, while the rate of frequency variation, *i.e.*, the number of times the frequency is changed between the minimum and maximum values per second, corresponds to the modulation frequency in amplitude modulation. Thus, if a 500-cycle sound wave is to be transmitted by frequency modulation of a 1,000,000-cycle carrier wave, this could be done by varying the transmitted frequency between 1,000,010 and 999,990 cycles, 500 times a second. If the pitch of the sound wave is increased to 1000 cycles, the carrier frequency would be varied between the same two limits 1000 times a second, while a sound wave of twice the intensity will be transmitted by varying the carrier frequency through twice the frequency range, *i.e.*, from 1,000,020 to 999,980 cycles in the above case.

Frequency-modulated waves can be readily produced by varying the capacity of the oscillator tuned circuit. The simplest way of doing this is to employ a small auxiliary condenser, one plate of which is a thin diaphragm that is vibrated by the voice currents in the same manner as is the diaphragm of a telephone receiver.[1] Reception is accomplished by converting the frequency-modulated wave into an amplitude-modulated wave by detuning the resonant circuits of the receiver so that the received frequency falls slightly to one side of the resonant point. As the frequency of the modulated wave varies, the response of the tuned circuit becomes alternately large and small, thus producing amplitude modulation.

Analysis of the Frequency-modulated Wave.—A superficial examination of frequency modulation might lead one to believe that intelligence could be transmitted in this way with an extremely narrow frequency band, since in the case cited above it appears that only 20 cycles band width is required to transmit the 500-cycle sound wave. This is not correct, however, because the variation in the frequency prevents the individual cycles from being exactly sinusoidal in shape. This is illustrated in Fig. 233*c*, where it is apparent that, since the changing frequency

[1] In the case of a high-power oscillator tube the diaphragm and its driving mechanism can be placed in an evacuated chamber, thereby enabling the arrangement to handle high voltages with small clearances between the plates.

Another very practical method makes use of the fact that the sending-end reactance of a transmission line one-eighth wave length long depends on the receiving-end resistance, and this resistance can be supplied by the plate circuit of a tube and varied by applying the signal to the grid. See Austin V. Eastman and Earl D. Scott, Transmission Lines as Frequency Modulators, *Proc. I.R.E.*, vol. 22, p. 878, July, 1934.

causes the time required to complete one-quarter cycle to differ from the time required by the next quarter cycle, the actual wave contains more than a single frequency. In fact, exact analysis shows that the frequency-modulated wave not only contains the same side-band frequencies as does the amplitude-modulated wave but also has higher order side bands that differ from the carrier frequency by integral multiples of the modulation frequency. Thus, when a carrier wave of frequency f_o is frequency modulated at a rate of f_s cycles per second, the resultant wave contains components having frequencies of f_o, $f_o + f_s$, $f_o - f_s$, $f_o + 2f_s$, $f_o - 2f_s$, $f_o + 3f_s$, $f_o - 3f_s$, etc.

The exact nature of a frequency-modulated wave can be determined by writing down the equation giving the instantaneous wave amplitude and then determining the frequency components contained in the result. The frequency-modulated wave can be readily shown to be expressible by the following equation[1]

$$i = I_m \sin(\omega t + \Phi + m_f \sin vt)$$

where ω and v are 2π times the carrier and audio frequencies, respectively, Φ is an arbitrary phase constant, and m_f (called the modulation index) is the ratio

$$\frac{\text{Variation in carrier frequency away from average carrier frequency}}{\text{Audio frequency}}$$

By making use of Bessel's functions it can be readily demonstrated that this wave consists of a carrier wave plus a series of side bands such as

[1] See Hans Roder, Amplitude, Phase, and Frequency Modulation, *Proc. I.R.E.*, vol. 19, p. 2145, December, 1931; Balth. van der Pol, Frequency Modulation, *Proc. I.R.E.*, vol. 18, p. 1194, July, 1930; John R. Carson, Notes on the Theory of Modulation, *Proc. I.R.E.*, vol. 10, p. 57, February, 1922.

The equation of the frequency-modulated wave is derived as follows: If the current is defined by the relation $i = A \sin \phi$, then the frequency at any instant is $(d\phi/dt)/2\pi$, and $\phi = \Phi + \int 2\pi f dt$, where Φ is an arbitrary phase constant. In the case of frequency modulation, the instantaneous frequency f is $f = f_o(1 + k_f \cos vt)$, where k_f is a constant. Substituting this in the integral giving ϕ, and denoting $(\omega t + \Phi)$ by $\omega_o t$, results in

$$i = A_o \sin(\omega_o t + m_f \sin vt)t$$

where

$$m_f = \frac{\text{variation in radio frequency away from the mean}}{\text{audio frequency}}$$

This can then be rewritten as

$$i = A_o\{I_o(m_f)\sin(\omega_o t) + I_1(m_f)[\sin(\omega_o + v)t - \sin(\omega_o - v)t] + I_2(m_f)[\sin(\omega_o + 2v)t + \sin(\omega_o - 2v)t] + \cdots\}$$

where $I_n(x)$ means the Bessel function of the first kind and the nth order.

described above. Since Bessel's functions are not familiar to most engineers, the results to which they lead have been plotted in Fig. 234, which shows the amplitude of the first-, second-, and third-order side-band components, *i.e.*, the side bands that differ from the carrier frequency by f_s, $2f_s$, and $3f_s$. These curves show that, when the modulation index m_f is less than 1, *i.e.*, when the range through which the frequency is varied is less than the audio frequency, the amplitude of the first-order side band is approximately proportional to the modulation index, while the higher order side bands are comparatively small.[1] When the modulation index exceeds unity, *i.e.*, when the range through which the radio frequency is varied is greater than the signal frequency, the second and other higher order components become of importance while the carrier amplitude drops rapidly and may even be zero. When the modulation index exceeds unity, there will be side-band components of appreciable magnitude extending on either side of the carrier up to the extreme limits between which the radio frequency is varied.

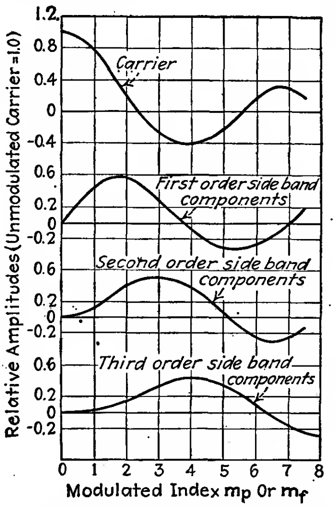


FIG. 234.—Amplitudes of frequency components of a frequency- or phase-modulated wave assuming that the amplitude of the unmodulated wave is 1.0. In the case of the side bands the amplitude shown is the amplitude of the individual side-band component and not of the pair of companion side bands taken together.

Frequency modulation is not particularly satisfactory as a means of transmitting intelligence. The frequency band is at least as great as that employed with amplitude modulation and is in general somewhat greater. Also the reception of frequency-modulated signals is not so simple a matter as the reception of amplitude-modulated waves.

Phase Modulation.—In phase modulation the intelligence is transmitted by varying the phase rather than the frequency of the radio wave. The equation of a phase-modulated wave is as follows

$$i = I \sin (\omega t + \Phi + m_p \sin vt)$$

[1] The question naturally arises as to why frequency modulation under such conditions is not the same as amplitude modulation, since the modulated waves contain substantially the same frequency components in the two cases. The difference lies in the fact that the carrier phase with respect to the side-band phase differs by 90°. Hence a frequency-modulated wave in which the second and still higher order side bands are negligible can be converted into an amplitude-modulated wave by shifting the carrier phase 90° with respect to the side bands.

where I is the amplitude of the wave, ω and v are 2π times the radio and audio frequencies, respectively, Φ is an arbitrary constant, and m_p the angle in radians through which the phase is displaced about the average phase. This equation is seen to be identical with that for the frequency-modulated wave, the only difference being in the interpretation of the modulation index m, which in the case of phase modulation depends only on the amplitude of the modulation and is independent of the frequency of the audio signal. It is hence apparent that the phase-modulated wave contains the same side-band components as does the frequency-modulated wave, and, if the modulation indexes in the two cases are the same, the relative amplitudes of these different components will also be the same. As long as the modulation index is less than unity (*i.e.*, phase shifts less than 57.3°), only the first-order side-band components are of appreciable magnitude, but each additional 57.3° of phase shift will add another pair of important side-band components. The relative amplitude of the carrier and the first three side bands can be obtained from Fig. 234 for any given modulation index.

Combinations of Phase, Amplitude, and Frequency Modulation.—Frequency and phase modulation are often combined with amplitude modulation as undesirable by-products. For example in the plate-modulated oscillator the plate-supply voltage of the oscillator tube is varied in accordance with the intelligence being transmitted, and, since the generated frequency depends more or less upon the plate voltage, the oscillations actually generated possess both frequency and amplitude modulation. For this reason modulated oscillators are practically never used in radio communication.

Combined phase and amplitude modulation can occur in a number of ways. Thus, if the tank circuit of a modulated amplifier or linear amplifier is not tuned exactly to resonance, there will be a phase shift that will vary with the modulation because the effective tube resistance (and hence the phase shift) varies with the amplitude modulation. Another source of combined amplitude and phase modulation is energy transfer between the carrier wave after modulation and the unmodulated exciting frequency. Such coupling causes a phase shift in the exciting voltage applied to the modulated amplifier, and, as the phase shift depends upon the amount of energy transfer, it will vary with the amplitude modulation.[1]

Phase and frequency modulation that occurs as a by-product of amplitude modulation is very undesirable in radio transmitters. This is because such modulation produces high-order side-band frequencies, which represent energy radiated upon adjacent frequency bands and

[1] For further information see W. A. Fitch, Phase Shift in Radio Transmitters, *Proc. I.R.E.*, vol. 20, p. 863, May, 1932.

which may interfere with other communications. Phase modulation is particularly bad in this respect because the modulation index m_p of phase modulation is independent of the frequency of modulation, whereas with frequency modulation the index m_f is inversely proportional to the modulation frequency and so tends to be low when the modulation frequency is high enough to make the second- and third-order side bands lie in adjacent channels.

Problems

1. The equation of a modulated wave is

$$e = 25(1 + 0.7 \cos 5000t - 0.3 \cos 10{,}000t) \sin 5 \times 10^6 t$$

a. What frequency components does the modulated wave consist of, and what is the amplitude of each?

b. Sketch the modulation envelope and evaluate the degree of modulation for the peaks and troughs.

2. Write the equation of a 100-volt carrier wave of 1000 kc when modulated 40 per cent at 400 cycles.

3. Calculate the total band width to which a radio receiver should respond for satisfactory reception of (*a*) ordinary broadcast signals, (*b*) perfect reproduction of speech and music, (*c*) telegraph code signals sent at 30 words per minute, (*d*) radio signals that represent radio extensions of wire telephone systems.

4. The Class C amplifier of the example in Sec. 61 (page 328) is to be completely modulated, using plate modulation with a transformer-coupled modulator such as shown in Fig. 215*c*.

a. Specify the undistorted audio power that the modulator must develop and the effective load impedance that the tube offers to the modulator.

b. Design a suitable modulator using commercial tubes in a push-pull circuit. The design includes the selection of suitable tubes, specification of operating conditions, and a statement as to the turn ratio of the output transformer.

5. *a.* In Prob. 4, obtain a linearity curve by calculating and plotting the output when the total plate voltage (*i.e.*, direct-current voltage plus modulating voltage) is 0, 0.293, 1.0, 1.707, and 2 times the direct-current voltage.

b. From these results calculate the second-, third-, and fourth-harmonic distortion.

6. *a.* Design a grid-modulated amplifier (100 per cent modulation) using the Type 800 tube of Fig. 75 ($\mu = 15$) for a plate-supply voltage of 1000 and permitting the grid to go a reasonable amount positive. In this design specify the proper voltages for the grid, the approximate load impedance, the expected carrier power, and the approximate grid current to be expected at the crest of the modulation cycle.

b. Calculate the maximum permissible internal impedance of the source of modulating voltage for a second-harmonic distortion not to exceed 5 per cent as a result of grid current [using Eq. (156)], and, from this and the audio voltage required, design the audio-modulating system.

7. The average power loss at the screen grid is much higher in a suppressor-grid modulated amplifier than the screen loss when the same tube is used as a Class C amplifier. Explain.

8. Explain the action of a van der Bijl modulated Class A amplifier by a mathematical analysis based on the equivalent circuit of Fig. 151 and the related analysis in Sec. 55, using only the first- and second-order effects.

9. Design a modulator that will completely modulate the modulated oscillator of Fig. 226. The design should include determination of undistorted power required, selection of suitable modulator tubes, tube operating conditions, load impedance which the oscillator offers to the modulator, and essential details of a suitable coupling arrangement between modulator and oscillator.

10. Sketch a series of oscillograms illustrating the action of the variable impedance modulator of Fig. 230*c*, and give an explanation of the operation based on these oscillograms.

11. By making reasonable assumptions as to plate efficiencies involved, estimate the total direct-current input power required by modulator and modulated tubes when the carrier power is 1000 watts and the modulation is zero and 100 per cent, for (*a*) plate-modulated Class C amplifier, Class A modulator; (*b*) plate-modulated Class C amplifier, Class B audio modulators; (*c*) grid-modulated amplifier; (*d*) suppressor-grid modulated amplifier.

12. Make a mathematical analysis of the balanced modulator of Fig. 231, demonstrating the action whereby the side bands appear in the output while the carrier does not, and also showing what current components flow through the battery.

13. A wave is frequency modulated at an audio rate of 5000 cycles per second. Plot the relative amplitude of the first-, second-, and third-order side-band components as a function of the amount the frequency is varied away from the carrier up to 15,000 cycles.

14. Using a table of Bessels functions, determine the values of modulation index for which the second-, third-, and fourth-order side-bands of a frequency- or phase-modulated wave have an amplitude that is 1 per cent of the amplitude of the unmodulated carrier.

CHAPTER X

VACUUM-TUBE DETECTORS

82. Detection of Radio Signals.—Detection is the process of reproducing the transmitted intelligence from the modulated radio wave. Since all systems of radio communication in practical use transmit intelligence by varying the amplitude of the radiated wave, the detection process, which is also sometimes spoken of as demodulation, must ordinarily produce currents that vary in accordance with the amplitude of the modulated radio signals. This is done by rectifying the modulated signal to obtain a pulsating direct current varying in magnitude in accordance with the original signal.

Rectification is obtained by applying the modulated radio wave to a circuit in which the current that flows is not proportional to the impressed voltage. During the history of radio communication many types of detectors have been employed, including mechanical, electrolytic, crystal, magnetic (utilizing hysteresis), and others, but all these have been displaced by the modern vacuum tube.

Detectors are commonly described by such expressions as *power*, *weak-signal*, *square-law*, and *linear*. A detector is a power or weak-signal detector according to whether it is intended to rectify large or small radio-frequency voltages, with carrier amplitudes of 1 or 2 volts being generally considered as the dividing point between strong and weak signals. A linear detector develops a rectified output proportional to the amplitude of the input voltage while a square-law detector develops an output proportional to the square of the amplitude. Weak-signal rectifiers are always of the square-law type, while power detectors are usually, but not necessarily, linear rectifiers.

An ideal detector reproduces in its output the exact intelligence modulated upon the radio wave. If the detector fails to do this, distortion results, which may be of several types. The detector output may include frequencies that were not contained in the original modulation, thus giving rise to amplitude distortion. The detector may also discriminate between modulation frequencies, giving an output that depends upon the modulation frequency, and may thus introduce frequency distortion. Finally a detector may reproduce the different components in the original modulation in altered phase relations, resulting in phase distortion.

83. The Diode Detector.—The two-electrode tube, or diode, represents one of the simplest and also one of the most widely used detectors. Such detectors take advantage of the fact that current flows through a diode only when the plate is positive, and make use of a circuit arrangement such as illustrated in Fig. 235*a*, where the resistance-condenser combination *RC* represents the load across which the rectified output voltage is developed. The mechanism of operation of a diode detector when a large signal voltage is applied can be understood with the aid of Fig. 235. At each positive peak of input voltage the plate is more positive than the cathode, causing a pulse of plate current to flow, which charges the condenser *C* to a potential only slightly less than the peak voltage of the input signal. Between peaks the plate of the diode is less positive than the cathode as a result of the negative charge on condenser *C*, so that no current flows through the tube. During this interval some of the charge upon the condenser leaks off through the resistance *R*, only to be replenished at the next peak of the modulated wave. When equilibrium exists, the voltage developed across the condenser *C* is enough less than the crest voltage of the signal peaks so that the difference between the two, which represents the voltage between the plate and cathode of the diode tube, is just sufficient to produce the plate-current pulses required to replenish the charge that leaks off the condenser between peaks. In the usual case the resistance *R* is much greater than the plate resistance of the diode, and this equilibrium consequently corresponds to a voltage across the condenser *C* which is only slightly less than the peak voltage of the radio-frequency signal. When this is true, the voltage across the resistance *R* is a measure of the peak amplitude of the wave. Hence, as the envelope varies with the modulation, the voltage across the resistance *R* will vary likewise, thereby reproducing the modulation envelope and recovering the intelligence modulated on the carrier wave.

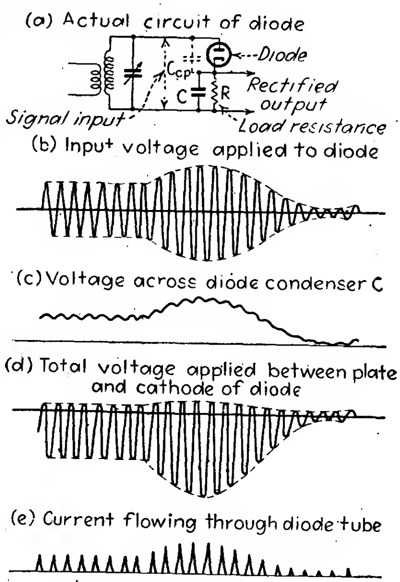


FIG. 235.—Circuit of simple diode detector, with oscillograms illustrating mechanism of operation.

The ratio of this output voltage developed across the resistance *R* to the voltage represented by the envelope of the modulated wave is the

detection efficiency, and can be made high by making the resistance R large compared with the diode plate resistance R_d. Under practical conditions the ratio R/R_d is of the order of 20 to 100, corresponding to efficiencies from 80 to 95 per cent.[1] In an ideal diode having a linear relationship between plate current and plate voltage the efficiency depends only upon the ratio R/R_d and is independent of signal voltage, but with the curved characteristic of actual tubes the efficiency increases somewhat as the signal becomes larger.[2]

The power that a diode detector absorbs from the input voltage can be represented by a resistance shunted across the diode input circuit. When the detection efficiency η is high, this equivalent input resistance is almost exactly $R/2\eta$, and so is slightly greater than one-half the load resistance R.[3] A high load resistance R therefore reduces the power absorbed by the detector because with a high load resistance the energy dissipated in the load is less in proportion to the voltage across it. The input resistance of a diode detector tends to be independent of the signal voltage in the case of an ideal tube having linear characteristics, but in the practical case it tends to drop somewhat with small applied voltages because of the curvature of the tube characteristics.

Distortion in Diode Detectors.—The simple diode detector of Fig. 235 may fail to reproduce faithfully the modulation envelope, either as a result of lack of proportionality between the signal voltage applied to the diode and the rectified voltage developed across the load resistance or as a result of inability of the voltage across the load resistance to follow the modulation envelope when the modulation frequency is high.

Lack of proportionality between the output voltage and modulation envelope is the result of curvature in the tube characteristic, making the

[1] The efficiency can be readily measured by applying a sine wave voltage to the detector input and determining the output voltage by observation of the resulting direct current flowing in the load resistance R. This measurement can be carried out using an audio voltage by making the condenser C large enough to be an effective by-pass across R for the frequency used.

[2] For further information on the detection efficiency and diode input resistance, see C. E. Kilgour and J. M. Glessner, Diode Detection Analysis, *Proc. I.R.E.*, vol. 21, p. 930, July, 1933.

[3] This is shown as follows: The diode current flows only when the signal voltage is at or near its crest value, as is clearly shown in Fig. 235. The power absorbed by the detector input is accordingly slightly less than the product of the crest signal voltage and the average diode current. Since the average current is equal to $\eta E/R$, where the output voltage across R is ηE and E is the crest value of signal voltage, one can write

$$\text{Power loss} = \frac{\eta E^2}{R} = \frac{(\text{effective signal})^2}{(R/2\eta)}$$

The denominator of this last term represents the equivalent input resistance to the signal, which is, accordingly, $R/2\eta$.

efficiency of rectification vary according to the envelope amplitude. Trouble from this source may be minimized by making the load resistance high compared with the diode plate resistance and by making the carrier amplitude applied to the diode reasonably large. The detection efficiency will then be high at all times even if not constant, and the rectified voltage developed across the load resistance will follow the modulation envelope quite closely. Under practical conditions when the load resistance and signal are sufficient to make the detection efficiency 80 per cent or higher, the distortion from this source is of the order of 2 per cent for a completely modulated wave. With small signals the distortion increases and ultimately reaches 25 per cent for a completely modulated wave when the signal voltage is a fraction of a volt.

The voltage across the condenser C can die away only as fast as the charge can leak off through the load resistance R, and this sets a limit to the maximum rate at which the modulation envelope can vary without introducing distortion. This is illustrated in Fig. 236, in which the conditions are such that, beginning with the point marked "O," the envelope of the input signal dies away faster than the voltage across the condenser can leak away through the load resistance. The result is a diagonal clipping off of the modulation envelope which produces both amplitude and frequency distortion in the output voltage. Trouble of this sort can be avoided by proportioning the resistance-condenser combination so that the rate of decay of voltage can follow the highest modulation frequency of importance. To achieve this it is necessary to satisfy the equation[1]

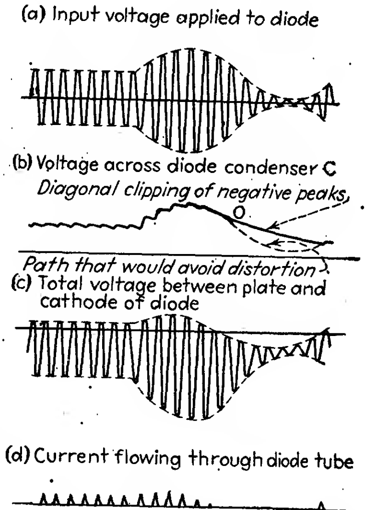


FIG. 236.—Operation of diode detector when output voltage cannot follow the modulation envelope, illustrating diagonal clipping of negative peaks.

$$\frac{X}{R} \geqslant \frac{m}{\sqrt{1 - m^2}} \tag{185a}$$

where X = reactance of the effective shunting capacity C at the modulation frequency in question and m = degree of modulation.

It is to be noted that for $m = 1$ (complete modulation) it is impossible for the output voltage to follow the modulation envelope at the troughs

[1] This equation can be derived as follows: The rate of change of voltage across a condenser C charged to a voltage E', caused by leakage through a resistance R, is

of the modulation cycle. However, if X/R is equal to 2 to 3, it is possible for the output voltage to follow the modulation envelope up to modulations of 0.9 to 0.95, and even with completely modulated waves the distortion will then be relatively small.[1]

Design of Simple Diode Detectors.—The desirable properties for a diode detector are high efficiency, high input resistance, and low distortion. Best results require a reasonably large carrier voltage (10 volts or more) together with a careful balance between the load resistance R and the shunting capacity C across the load.

The capacity C should be as small as possible in order to have a high reactance X to the modulation frequency, but at the same time must be at least five to ten times the plate-cathode interelectrode capacity of the

---

$-E'/RC$. If the equation of the envelope of the modulated signal wave is

$$e = E(1 + m \cos 2\pi ft)$$

then the rate of change of envelope is

$$\frac{de}{dt} = -mE2\pi f \sin 2\pi ft$$

and at the time $t = t_o$ the envelope magnitude e_o is

$$e_o = E(1 + m \cos 2\pi ft_o)$$

The grid condenser rate of discharge can follow the rate of change of modulation envelope at any time t_o of the modulation cycle provided $-E'/RC \leqslant de/dt$, where de/dt is evaluated at $t = t_o$ and $E' = e_o$, or when $-mE2\pi f \sin 2\pi ft_o \geqslant -E(1 + m \cos 2\pi ft_o)/RC$. This can be reduced to

$$\frac{X}{R} = \frac{1}{2\pi fCR} \geqslant \frac{m \sin 2\pi ft_o}{(1 + m \cos 2\pi ft_o)}$$

The point on the modulation cycle where it is most difficult for the grid-condenser charge to keep up with the envelope change is at a time t_o that makes the right-hand member of this last equation maximum, which is when $\cos 2\pi ft_o = -m$. The maximum value of the right-hand member of the equation is therefore $m/\sqrt{1 - m^2}$, and the charge on the grid condenser can decrease at least as fast as the modulation envelope changes when $X/R \geqslant m/\sqrt{1 - m^2}$.

This derivation and the relation expressed by Eq. (185)are due to F. E. Terman and N. R. Morgan, Some Properties of Grid Leak Power Detection, *Proc. I.R.E.*, vol. 18, p. 2160, December, 1930. A slightly different derivation is given by E. L. Chaffee, "The Theory of Thermionic Vacuum Tubes," p. 578, and leads to a $\eta\beta$ in the denominator of the right-hand term, but this is not quite so accurate, as it fails to take into account the fact that the current through the diode is zero at the critical part of the cycle.

[1] It is to be noted that Eq. (185*a*) can be rewritten as

$$m \leqslant \frac{\dfrac{XR}{\sqrt{X^2 + R^2}}}{R} = \frac{\left\{\begin{array}{l}\text{Impedance of diode load to the}\\ \text{modulation frequency}\end{array}\right.}{\left\{\begin{array}{l}\text{Resistance of the diode load}\\ \text{to direct current}\end{array}\right.} \qquad (185b)$$

The significance of this relation is discussed in detail later in this section.

diode in order that most of the signal voltage available will be used up as drop across the diode tube, and only a small part of the signal voltage will appear as drop across the diode condenser C. The values of diode capacity commonly employed are in the range 50 to 100 $\mu\mu$f. This includes stray capacities and the plate-cathode capacity of the diode, since these are effectively in shunt with C insofar as the modulation-frequency voltage developed across the load resistance is concerned. In some cases, particularly when the cathode is grounded, the stray capacity from the tuned input circuit to ground is sufficient to furnish all the shunting capacity required across the load resistance.

With the capacity C fixed by these considerations, the load resistance R is assigned the highest value that is permissible in view of Eq. (185a). A high resistance increases the detector efficiency and makes for high

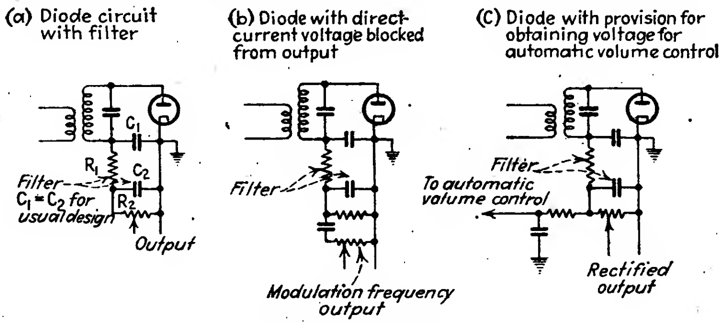


FIG. 237.—Diode detector circuits

input resistance to the radio-frequency signal, but at the same time tends to cause distortion at the higher modulation frequencies. The practical compromise is usually a value of resistance such that at the highest modulation frequencies of importance X/R is from 1 to 3. Values of R of the order of 250,000 ohms are commonly used in the reception of ordinary broadcast signals.

In practical detectors it is generally found necessary to provide a filter such as shown in Fig. 237a in order to prevent any radio-frequency voltage from reaching the output. In this filter the condensers C_1 and C_2 are commonly equal, while R_2 may be several times R_1. The total load resistance R with filter is $R_1 + R_2$, while the effective diode capacity C is between C_1 and $C_1 + C_2$. A filter reduces the output by the factor $R_2/(R_1 + R_2)$ as a result of the voltage drop in R_1.

Diode Rectifiers with Complex Load Impedance.[1]—It is commonly desirable to block off from the output the direct-current voltage developed

[1] The material in this section is based largely upon studies made by H. A. Wheeler under the auspices of the Hazeltine Corporation, and made available to the author by that organization and Mr. Wheeler.

across the diode load resistance, by the use of a resistance and condenser as illustrated in Fig. 237*b*. It is also common practice to use the direct-current voltage developed across the diode load resistance for automatic volume control by means of an arrangement such as illustrated in Fig. 237*c* for separating this direct voltage from the alternating modulation-frequency voltage. These modifications cause the impedance that the load offers to the modulation-frequency component of the output voltage to differ from and in general be less than the resistance that the diode load circuit offers to the rectified direct current. The result is a modification of the properties of the diode detector in several important respects.

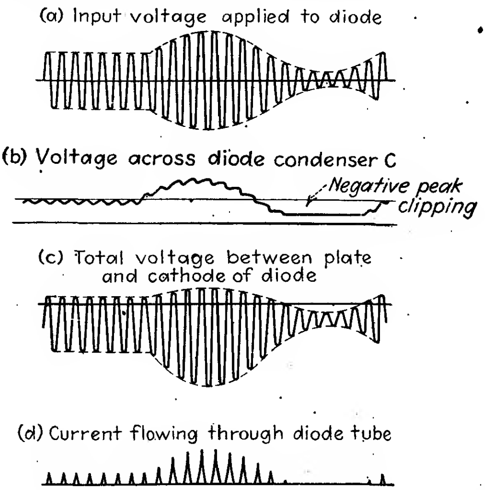


FIG. 238.—Operation of diode detector when the diode load circuit offers a lower resistance impedance to the modulation frequency than to direct currents, showing the clipping of the negative peaks that results when the degree of modulation is high.

When the impedance offered to the direct-current component of the rectified current differs from the load impedance to the alternating current of modulation frequency, it is found that, when the degree of modulation is high, the negative peaks are clipped off as illustrated in Fig. 238, with resulting amplitude distortion.[1] The reason for this can be understood by the following explanation: An unmodulated carrier wave applied to the diode detector produces a direct-current component of the rectified current such that the voltage which this current develops in flowing through the direct-current resistance offered by the diode load circuit is just slightly less than the crest carrier voltage. When the carrier is modulated, the rectified current through the diode varies, giving the equivalent of an alternating current of modulation frequency superimposed upon the direct current. The magnitude of this alternating

[1] This was first observed by Kilgour and Glessner, *loc. cit.*

current is such that the voltage drop which it produces in flowing through the alternating-current impedance that the load offers is just less than the crest alternating voltage contained in the modulation envelope. It is apparent that, if the impedance which the diode load offers to the alternating current is less than the impedance offered to direct current, then the alternating current that flows through the diode will be larger than if the load impedance were the same to all components of the rectified current. Now the maximum possible amplitude that the alternating current can have without being distorted is a peak value equal to the direct current produced by the carrier, because, when this is superimposed upon the d-c current, the instantaneous current through the diode then reaches zero during each modulation cycle. Hence the maximum degree of modulation that can be rectified without distortion is given by the equation[1]

$$\left.\begin{matrix}\text{Maximum allowable}\\ \text{degree of modulation}\\ m \text{ for distortionless}\\ \text{rectification}\end{matrix}\right\} = \frac{\left\{\begin{matrix}\text{Magnitude of the impedance of diode}\\ \text{load to modulation frequency}\end{matrix}\right.}{\left\{\begin{matrix}\text{Resistance of the diode load}\\ \text{circuit to direct current}\end{matrix}\right.} \qquad (186a)$$

Whenever the degree of modulation exceeds the value given by Eq. (186a), the negative peaks will be clipped off.[2]

When the diode load circuit offers different impedances to alternating and direct currents, the input impedance that the diode offers to the source of exciting voltage is not the same for the side-band frequencies as for the carrier. The input resistance to the carrier as already derived is $R/2\eta$, where R is the load resistance offered to direct current and η is the efficiency of detection. The input resistance offered to the side-band frequencies is $Z_m/2\eta$, where Z_m is the impedance that the diode load offers to the modulation frequency. Hence, when the impedance of the diode load circuit is less for alternating than for direct currents, the side-band components of the input modulated wave tend to be reduced. This causes the degree of modulation of the wave applied to the diode to be less than the degree of modulation of the actual signal by an amount that depends upon the ratio Z_m/R, and upon the output imped-

[1] It will be noted that this equation is a general case of Eq. (185a), as it must be since the latter is for an analogous situation in which the load impedance offered to the high modulation frequencies is less than the load resistance offered to direct current as a result of the shunting effect of the diode condenser C.

[2] When the amount of clipping is small, the total r.m.s. distortion that results is roughly:

$$\left.\begin{matrix}\text{Approximate}\\ \text{r.m.s. distortion}\end{matrix}\right\} = \frac{[\text{Actual modulation}] - [\text{modulation allowed by Eq. (186}a)]}{2\ (\text{actual modulation})} \qquad (186b)$$

ance Z_s and Z_s' of the source of exciting voltage to carrier and side-band frequencies, respectively, according to the equation[1]

$$\frac{\left.\begin{matrix}\text{Degree of modulation of}\\ \text{diode input voltage}\end{matrix}\right\}}{\left.\begin{matrix}\text{Actual degree of modulation}\\ \text{of original signal}\end{matrix}\right\}} = \frac{|Z_m| \times \left|\left(Z_s + \dfrac{R}{2\eta}\right)\right|}{R\left|\left(Z_s' + \dfrac{Z_m}{2\eta}\right)\right|} \tag{187}$$

In Eq. (187) the angle assigned to Z_m is leading for the upper and lagging for the lower sideband. When the output impedance Z_s of the source is high and the ratio Z_m/R departs appreciably from unity, the resulting reduction in degree of modulation is considerable. This makes it much easier to satisfy Eq. (186) and thus tends to reduce the amplitude distortion that would otherwise occur when the signal as transmitted has a high degree of modulation.

Examination of the diode load circuit shows that the impedance which is offered to the modulation frequency decreases at very high modulation frequencies as a result of the shunting condenser C. At high modulation frequencies having a high degree of modulation this tends to cause a diagonal clipping of the negative peaks as illustrated in Fig. 236. However, if the input circuit tends to discriminate against the higher modulation frequencies, this reduces the degree of modulation of the voltage actually applied to the diode. Where this reduction varies with modulation frequency in exactly the same way as the diode load impedance falls off with frequency, diagonal clipping of the output wave is avoided. The higher modulation frequencies are then discriminated against by being reproduced at lower than normal volume, but the rectification process produces no harmonics.[2]

A typical design of a diode rectifier arranged to supply a direct-current voltage for automatic-volume-control purposes, and at the same time develop a modulation-frequency output free of superimposed direct

[1] This is derived as follows: To the carrier the diode offers a load resistance $R/2\eta$, so that the presence of the diode reduces the carrier voltage across the tuned input circuit by the factor $|(R/2\eta)/(Z_s + R/2\eta)|$. Since the input impedance to the side bands is $Z_m/2\eta$, the presence of the diode reduces these by the factor $\left|\left(\dfrac{Z_m}{2\eta}\right)\right| / \left|\left(Z_s' + \dfrac{Z_m}{2\eta}\right)\right|$. The ratio of these two reduction factors gives the factor by which the modulation is altered, and leads at once to Eq. (187).

[2] Thus, when the signal is completely modulated at a frequency such that the tuned input circuit gives 70.7 per cent response to the side-band frequencies, the degree of modulation of the voltage actually applied to the diode is only 70.7 per cent, and according to Eq. (186a) the required ratio Z_m/R is 0.707. It will be noted that this corresponds to $X/R = 1$ in Eq. (185a), which makes the diode load impedance 0.707 of its value to direct current.

current, is given in Fig. 239. Here the combination R_1C_2 is a filter that prevents carrier-frequency voltages developed across C_1 from reaching the modulation-frequency output. The resistance-condenser combination R_4C_4 across the resistance R_2 is a filter for delivering modulation-frequency output free from superimposed direct current, while the combination R_3C_3 is a filter for delivering to the automatic volume control direct-current voltage proportional to the carrier and free of modulation. In shunting these last two resistance-condenser combinations across R_2 instead of across the condenser C_1, some of the output voltage is thrown away as drop in resistance R_1. This is necessary, however, in order to provide radio-frequency filtering and in order that the effective load impedances to alternating and direct currents will not be too widely

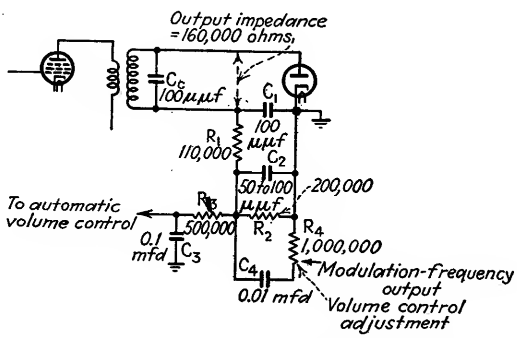


FIG. 239.—Typical design of diode detector in which provision is made for an automatic-volume-control voltage and for obtaining the modulation-frequency output free of direct-current voltage.

different. With the proportions shown in Fig. 239, $Z_m/R = 0.76$ at the lower modulation frequencies, and the degree of modulation of the diode input voltage at which peak clipping begins is 0.76. However, because of the fact that the input resistance to the side-band frequencies is lower than to the carrier, substitution in Eq. (187) shows that, for the input circuit of Fig. 239 and assuming the detection efficiency is 90 per cent, the original radio signal must be modulated 88 per cent to make the degree of modulation at the diode terminals 76 per cent. With a completely modulated signal voltage there will be some negative peak clipping, but the amount will be relatively small, representing a distortion of approximately 5 per cent. Harmonic distortion at the higher side-band frequencies has been minimized in the design of Fig. 239 by making $C_c = C_1$ and proportioning so the input resistance of the diode approximately equals the output resistance of the tuned exciting circuit. The impedance of the diode load circuit then falls off with modulation frequency in the same way as does the output of the exciting circuit, thereby

satisfying the requirement for high-frequency reproduction without harmonic distortion.

Tubes for Diode Detectors.—The tube used as a diode detector should if possible have a low plate resistance. The diodes employed in actual practice include small especially designed two-electrode tubes, diode sections built into triode and pentode tubes, and also ordinary triodes converted into diodes by connecting the grid and plate electrodes together.

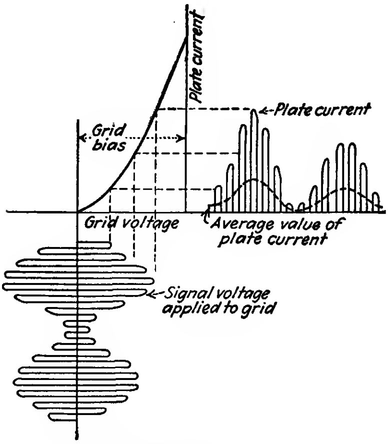


FIG. 240.—Details of action taking place in plate rectifier, showing how a modulated wave applied to the grid adjusted to cut-off will cause plate-current impulses having an average value that varies in accordance with the modulation envelope.

84. Plate Power Detectors.—In the plate detector, also sometimes called bias detector and anode detector, advantage is taken of the fact that, when an amplifier tube is biased approximately to cut-off and a relatively large alternating voltage is applied to the grid, there will be a pulse of plate current during each positive half cycle of the applied voltage and no plate current during the negative half cycles, as illustrated in Fig. 240. The amplitude of these pulses of plate current is proportional to the alternating voltage applied to the grid so that, if this input is a modulated wave, the average value of the plate current will vary in accordance with the modulation envelope and thereby reproduce the intelligence contained in the transmitted signal. A useful output is obtained by placing in the plate circuit of the tube an impedance to the

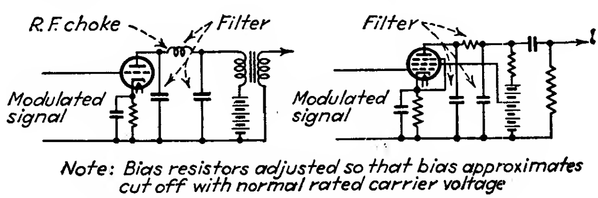


FIG. 241.—Typical plate-detector circuits.

modulation frequency. Typical circuit arrangements suitable for anode detection are illustrated in Fig. 241. When triode tubes are employed, either resistance or transformer coupling is satisfactory, while with pentode detector tubes resistance coupling is generally preferred. Some form of filter such as illustrated in Fig. 241 is always necessary to prevent radio-frequency currents from reaching the output.

Analysis of Plate Detectors.[1]—The quantitative relations that determine the modulation-frequency output of a plate detector can be expressed in terms of the equivalent circuit of Fig. 242*a*. Here the plate circuit of the detector tube has been replaced by a generator developing a modulation-frequency voltage of E_R and having an internal resistance R_d.[2] It will be noted that this equivalent circuit is of exactly the same form as the equivalent circuit of the amplifier (see Sec. 43 and Fig. 91), and so can be made use of in the same way in calculating the output voltage, frequency-response characteristic, proper load imped-

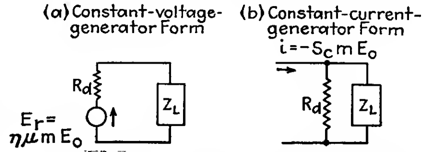


FIG. 242.—Equivalent plate circuits of plate detector. Note the similarity to the equivalent amplifier circuits of Fig. 91.

[1] This treatment of anode power detection is a modification of the method of analysis originated by Stuart Ballantine, Detection of High Signal Voltages—Plate Rectification with a High Vacuum Triode, *Proc. I.R.E.*, vol. 17, p. 1153, July, 1929.

The analysis of the anode detector is carried out by a method differing from that used with the diode because of the fact that under practical circumstances, particularly when dealing with pentode tubes, the diode method of approach is quite cumbersome and does not enable the frequency response to be easily determined.

[2] The proof for this is derived as follows: When a carrier voltage E_0 modulated to a degree m is applied to the grid of the tube, the amplitude of the alternating input voltage varies sinusoidally about E_0 by an amount mE_0. This variation in the signal voltage produces a variation in the rectified plate current of ΔI_p.

If the degree of modulation is not too great, then to a first approximation one can write

$$\Delta I_p = \frac{\partial I_p}{\partial E_s} mE_0 + \frac{\partial I_p}{\partial E_p}\Delta E_p$$

where $\partial I_p/\partial E_p$ represents the rate of change of rectified plate current with plate voltage as evaluated in the presence of an unmodulated carrier voltage E_0, ΔE_p represents the change of plate voltage produced by the rectified current ΔI_p flowing through the load impedance Z_L in the plate circuit, and $\partial I_p/\partial E_s$ is the rate of change of rectified plate current with carrier voltage as taken at constant plate voltage and evaluated about a carrier voltage E_0. Substituting $\Delta E_p = -Z_L \Delta I_p$ in the equation gives

$$\Delta I_p = \frac{\partial I_p}{\partial E_s} mE_0 - Z_L \Delta I_p \frac{\partial I_p}{\partial E_p}$$

Solving this for ΔI_p results in

$$\Delta I_p = \frac{mE_0\left(\dfrac{\partial I_p/\partial E_s}{\partial I_p/\partial E_p}\right)}{\left(Z_L + \dfrac{1}{\partial I_p/\partial E_p}\right)}$$

This equation then represents a voltage

$$mE_0 \frac{\partial I_p/\partial E_s}{\partial I_p/\partial E_p}$$

acting in a circuit having an impedance comprising a load impedance Z_L in series with a resistance $\partial E_p/\partial I_p$.

ance, etc. When dealing with anode detectors employing pentode tubes, it is convenient to transform this equivalent circuit into the constant-current form shown in Fig. 242*b*, which is analagous to the constant-current form of equivalent amplifier circuit shown in Fig. 91*c*.

The generator voltage E_R that can be considered as acting in the equivalent circuits of Fig. 242 is given by the expression

$$E_R = D\mu m E_0 \tag{188}$$

where

μ = amplification factor of tube
m = degree of modulation of carrier
E_0 = carrier voltage
$D = \dfrac{(\partial I_p/\partial E)_{E=E_0}}{\mu(\partial I_p/\partial E_p)_{E=E_0}}$ = efficiency of conversion
E = alternating voltage applied to grid
I_p = direct-current component of plate current
E_p = direct-current voltage applied to the plate.

The maximum possible value that the efficiency of conversion D can have is unity, and practical values are commonly of the order of 0.8 if the applied carrier voltage is reasonably large. The exact value depends upon the carrier amplitude and upon the direct-current voltages actually applied to the grid and plate, but is independent of the load impedance connected in series with the plate circuit except insofar as this load impedance may alter the direct-current voltage at the plate.[1]

The detector plate resistance R_d is the dynamic resistance of the plate circuit when measured at the operating point in the presence of an applied carrier. That is

$$R_d = \left.\frac{\partial E_p}{\partial I_p}\right|_{E = E_0} \tag{189}$$

The value of the detection plate resistance R_d depends upon the direct-current voltages actually applied to the grid and plate and upon the carrier amplitude, but is independent of the load impedance placed in series with the plate circuit except insofar as the direct-current plate voltage is affected by the load. With large signal voltages and a bias

[1] The efficiency of detection D can be measured by applying to the grid of the detector an alternating voltage (of any convenient frequency) equal to the carrier voltage E_0 for which D is desired and by noting the d-c plate current. The crest value of this alternating grid voltage is then increased by a small increment ΔE_s, after which the plate voltage is altered by any amount ΔE_b such that the d-c plate current is the same value as before the addition of the grid voltage increment. The ratio $\Delta E_b/\Delta E_s$ is then μD.

that approximates cut-off, the effective plate resistance of the detector usually exceeds twice the plate resistance obtained for amplifier operation under normal conditions.

The quantity $D\mu/R_d = S_c$ has the dimension of a conductance, and can be termed the conversion transconductance (or the conversion mutual conductance). This ratio is of particular significance in the constant-current form of the equivalent plate circuit of the anode detector as shown in Fig. 242*b*, and is of especial usefulness in connection with anode detectors employing pentode tubes. The value of the conversion transconductance S_c for pentode tubes with a large carrier voltage and a bias approximating cutoff is normally about 0.3 to 0.4 of the mutual conductance for normal amplifier operation.

The equivalent circuits of the anode detector illustrated in Fig. 242 are exactly correct when the degree of modulation is small, provided the efficiency of detection, conversion transconductance, and detection plate resistance are evaluated for the direct-current voltages actually applied to the grid and plate electrodes in the presence of the unmodulated carrier. The circuits are also correct for high degrees of modulation provided the detector is reasonably linear. If the grid of the tube is not driven positive at the crest of the modulation cycle, the plate detector will generally be substantially linear except when the carrier becomes quite small. The detector then becomes a square-law device (see next section).

Design of Anode Detectors.—The load impedance placed in series with the plate of the anode power detector is designed exactly as in the case of voltage amplifiers, but using the equivalent detector plate circuit of Fig. 242 instead of the equivalent amplifier circuit of Fig. 91. The only essential difference in the circuit arrangements is that in the detector it is necessary to include in the coupling arrangements a by-pass condenser and filter as shown in Fig. 241 to prevent radio-frequency voltages from reaching the output, while providing a low impedance path in the plate circuit to radio-frequency currents. The condensers in this filter must be small, for they are effectively in shunt with the load impedance and so tend to reduce the high-frequency response, particularly when resistance coupling is used.

The tubes used for audio-frequency voltage amplification are also suitable for anode power detection. Both triode and pentode tubes are satisfactory, although the latter are preferred because of their greater gain. The grid bias should be approximately cut-off or slightly less, and so depends upon the screen potential in pentodes and upon the plate potential in triodes. When a bias resistor is used, its value should be chosen so that the necessary bias is obtained *in the presence of the rated carrier input voltage.*

It is generally desirable to avoid driving the grid positive as this causes energy to be absorbed at the peaks of the modulation cycle and therefore tends to introduce distortion. If the grid is not driven positive, the largest carrier voltage that can be applied has a crest value half the grid bias. This fixes the power-handling capacity of the tube, which is hence larger when the screen or plate potential is greater. With a given degree of modulation the detector will be more linear the larger the

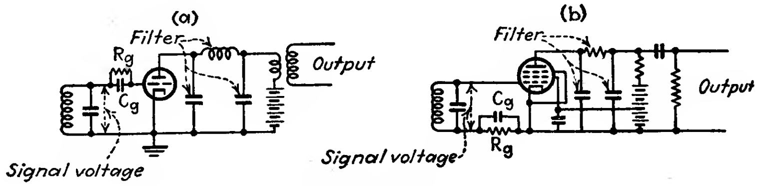


FIG. 243.—Circuits of grid-leak power detector.

carrier voltage, up to the point where the grid becomes positive at the crest of the modulation cycle, while with a given carrier voltage the amplitude distortion increases as the degree of modulation approaches 100 per cent.

85. Miscellaneous. *Grid-leak Power Detector.*[1]—Typical circuit arrangements for grid-leak power detection are shown in Fig. 243, where the grid-leak resistance R_g and grid-leak condenser C_g form an impedance that corresponds to the load impedance RC of the diode detector (see Fig. 235*a*). The grid circuit, comprising the source of exciting voltage, the leak-condenser combination R_gC_g, and the grid-cathode of the tube, functions as a diode rectifier exactly as illustrated in Figs. 235, 236, and 238, with the grid electrode performing the same function as the plate of the diode. The mechanism of operation is illustrated in Fig. 244. The voltage existing between the grid and cathode of the tube consists of the radio-frequency signal voltage plus the voltage drop developed cross the load impedance R_gC_g by the rectified grid current. Since this latter potential is applied to the grid of the tube, it is amplified by the use of a suitable plate-coupling arrangement. The

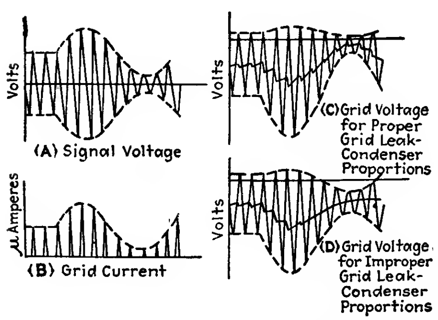


FIG. 244.—Details of action taking place in grid circuit of grid-leak power detector of Fig. 243 when a modulated radio-frequency voltage is applied to the detector.

[1] A somewhat more extended discussion of grid power detection is to be found in, F. E. Terman and N. R. Morgan, Some Properties of Grid-leak Power Detection, *Proc. I.R.E.*, vol. 18, p. 2160, December, 1930.

amplified modulation-frequency component can then be separated and utilized as desired.

The grid-leak power detector corresponds in many respects to a diode rectifier plus one stage of amplification. In particular, the leak-condenser combination R_gC_g is proportioned according to the same factors effective in determining the proportions of R and C in the diode (see Sec. 83), and the input resistance, requirements for avoiding frequency distortion, and efficiency of detection are all determined by the same factors in both diode and grid-leak power detectors. The usual proportions for the leak-condenser combination is therefore a grid-leak resistance R_g normally between 100,000 and 500,000 ohms and a grid-condenser capacity C_g of the order of 50 to 100 $\mu\mu$f.

The maximum signal voltage that can be applied to a grid-leak power detector without producing distortion is limited by the fact that the tube must amplify without distortion the total voltage applied between grid and cathode if the modulation-frequency voltage developed across the load impedance R_gC_g is to be properly reproduced. Distortion is most likely to occur at the crest of the modulation cycle, since this is when the maximum voltage is applied to the grid. With a completely modulated wave of carrier amplitude E_0 and a detection efficiency of η, the exciting voltage has a crest amplitude $2E_0$, and the total voltage across R_gC_g has a maximum of $2\eta E_0$, where E_0 is the carrier amplitude. This makes the maximum negative instantaneous voltage applied to the grid equal to $2(1 + \eta)E_0$, which must be amplified in the plate circuit with negligible amplitude distortion. It is apparent that the amount of voltage which can be handled is proportional to the direct-current plate potential at the operating point. Since the permissible plate potential is limited by the fact that, when no carrier voltage is applied, the grid bias is zero, the power-handling capacity of a grid-leak power detector is appreciably less than the power capacity of the same tube when operated as a Class A amplifier. With ordinary tubes the maximum carrier voltage that can be handled without overloading is approximately one-third to one-half of the signal voltage that can be handled by the same tube acting as an audio-frequency amplifier at the same plate voltage, and the modulation-frequency output voltage obtainable is roughly 0.3 to 0.4 of the output obtainable from the corresponding amplifier.

The Wunderlich Tube.[1]—The Wunderlich coplanar-grid tube was originally developed as a grid-leak power detector for a circuit such as illustrated in Fig. 245. Here each individual grid functions in cooperation with the cathode to rectify the voltage applied between it and the

[1] For additional information on the theory of the Wunderlich tube, see F. E. Terman, Further Description of the Wunderlich Tube, *Radio Eng.*, vol. 12, p. 25, May, 1932.

center tap of the input circuit, with the load impedance R_gC_g common to both grids. It will be observed that this arrangement is essentially the same as the full-wave center-tapped rectifier circuit illustrated in Fig. 266*b* and commonly used to supply plate power for radio receivers. The voltage developed across the load R_gC_g by the rectified grid current is applied to the two grids in parallel, and so is amplified in the plate circuit by ordinary amplifier action. The radio-frequency input voltage is applied to the two grids in phase opposition, so that, as far as radio-frequency voltages are concerned, the two grids neutralize each other's effect on the plate current, and the plate circuit does not carry radio-frequency currents. The result is that the allowable modulation-frequency voltage that can be developed by a Wunderlich tube is substantially the same as the audio-frequency voltage that can be developed by a corresponding triode tube at the same plate potential. The Wunderlich tube hence develops approximately three times as much output voltage as the ordinary grid-leak power detector before overloading. The considerations determinining the grid-leak and grid-condenser proportions are the same in the Wunderlich tube as in a diode detector except that the input resistance from grid to grid is approximately $2R/\eta$ as a result of the center-tapped input arrangement.

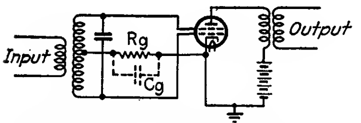


FIG. 245.—Circuit of Wunderlich detector tube.

Square-law Detectors.—The term *square law* is applied to any detector in which the rectified direct-current output is proportional to the square of the effective value of applied signal voltage. The simplest form of square-law detector is an amplifier adjusted as shown in Fig. 246 so that the operating point is on the curved part of the grid-voltage plate-current characteristic curve. When a signal voltage is applied to such an amplifier, the positive half cycles are amplified more than the negative half cycles, causing an increase in the d-c plate current and also introducing frequency components in the output current not present in the applied signal.

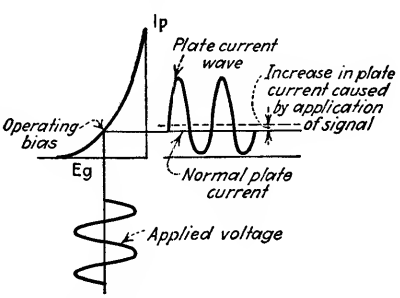


FIG. 246.—Vacuum tube adjusted to operate as a square-law device.

When the instantaneous plate current never reaches zero, the situation can be analyzed by the power-series method discussed in Sec. 55, from which the behavior can be expressed in terms of the generalized

equivalent circuit of Fig. 151. If the signal voltage is not too large, the limited portion of the tube characteristic made use of can be approximated by a section of a parabola (*i.e.*, only the first two terms in Eq. (126) need be considered), and the behavior then follows that of a true square-law device. Under such conditions the output can be divided into two components: first, an amplified reproduction of the input signal; second, direct-current, harmonic, and combination frequency terms. It is this second component of the plate current that is responsible for the detector action of a square-law device.

These second-order effects are discussed in Sec. 55 and so will be summarized here only very briefly as follows: The contributions of the second-order action consist of a d-c rectified current proportional to the square of the effective value of the applied signal, together with second harmonics of each frequency component of the applied signal and sum and difference frequencies formed by every possible combination of frequencies contained in the input signal. The amplitude of each second-harmonic component is proportional to the square of the amplitude of the corresponding fundamental part of the signal, while each combination frequency has an amplitude proportional to the product of the amplitudes of the individual components involved. Thus, when the signal consists of sine wave of amplitude E_1 and frequency f_1, the output will consist, in addition to the amplified sine wave, of a direct-current term having an amplitude proportional to E_1^2 and a second-harmonic term of frequency $2f_1$ and amplitude proportional to E_1^2. When the input signal consists of the sum of two sine waves having amplitudes E_1 and E_2 and frequencies f_1 and f_2, the output will contain, in addition to the amplified signal, second-order contributions consisting of a direct-current term proportional to $\frac{E_1^2}{2} + \frac{E_2^2}{2}$, second-harmonic terms of frequencies $2f_1$ and $2f_2$ with amplitudes proportional to E_1^2 and E_2^2, respectively, sum and difference frequency terms having frequencies $f_1 + f_2$ and $f_1 - f_2$, respectively, and amplitudes proportional to E_1E_2. When a signal is a modulated wave, the fact that the rectified output is proportional to the square of the amplitude of the signal causes the output to be a distorted reproduction of the modulation envelope. With sinusoidal modulation the distortion is all second harmonic and has a percentage $100 \times m/4$. For purposes of analysis the modulated wave can be thought of as three frequency components, the carrier and two side bands. The output then contains the rectified direct-current term, second harmonics of each of the three components, two difference-frequency terms formed by the carrier combining with the two side bands to give the desired modulation-frequency output, and a second harmonic of the modulation frequency arising from the difference frequency formed by the two side-band components.

A square-law detector is of importance for a number of reasons. In the first place, any detector becomes a square-law device when the applied signal is quite small. The square-law detector is also shown in Sec. 87 to give distortionless detection of heterodyne signals. Moreover, square-law detectors also have a simple and exact law of behavior, and do not generate high-order harmonics of the carrier frequency by producing current pulses representing chopped-off sections of sine waves. The square-law detector is not particularly efficient, but it finds considerable use in special circumstances, as, for example, in vacuum-tube voltmeters and other laboratory measuring equipment.

The usual form of square-law detector is an amplifier biased so that the operating point is on the curved part of the plate-current characteristic, as shown in Fig. 246. A modification of this arrangement is to bias the tube approximately to cut-off and to apply a moderately large signal voltage. This gives a square-law response to the positive half cycles of the applied voltage with suppression of the negative half cycles, resulting in a half-wave square-law device sometimes used in vacuum-tube voltmeters.

Square-law detectors making use of grid-leak grid-condenser arrangements were widely used in the early broadcast receivers. The circuit is the same as for grid-leak power detection, but the grid-leak resistance and grid-condenser capacity are proportioned differently, values of 1 to 5 megohms and 0.00025 μf being common. Such a weak-signal grid-leak detector makes use of the curved relationship existing between grid current and grid voltage in the vicinity of zero grid potential to produce a rectified grid current which in flowing through the leak-condenser combination R_gC_g develops a modulation-frequency voltage that is amplified in the plate circuit. Such an arrangement is the most sensitive of all practical weak-signal detectors, but is now no longer used. The reader who wishes to study the weak-signal grid-leak detector is referred to the extensive literature upon the subject.[1]

[1] The following is a list of references selected from the extensive literature on the detection of weak signals: Stuart Ballantine, Detection by Grid Rectification with the High-vacuum Triode, *Proc. I.R.E.*, vol. 16, p. 593, May, 1928; John R. Carson, The Equivalent Circuit of the Vacuum-tube Modulator, *Proc. I.R.E.*, vol. 9, p. 243, June, 1921; E. L. Chaffee and G. H. Browning, A Theoretical and Experimental Investigation of Detection for Small Signals, *Proc. I.R.E.*, vol. 15, p. 113, February, 1927; F. M. Colebrook, The Rectification of Small Radio-frequency Potential Differences by Means of Triode Valves, *Exp. Wireless and Wireless Eng.*, vol. 2, p. 946, December, 1925; F. E. Terman, Some Principles of Grid-leak Grid-condenser Detection, *Proc. I.R.E.*, vol. 16, p. 1384, October, 1928; F. E. Terman and T. M. Googin, Detection Characteristics of Three-element Vacuum Tubes, *Proc. I.R.E.*, vol. 17, p. 149, January, 1929.

Vacuum-tube Voltmeters.[1]—A detector can be used as a voltmeter by making use of the rectified d-c current to measure the voltage applied to the detector. The resulting vacuum-tube voltmeter is one of the most useful measuring devices available for audio and radio frequencies. It consumes little or no power from the voltage being measured, and when properly designed it can be calibrated at a low frequency, such as 60 cycles, and used at any radio or audio frequency up to ultra-high frequencies.[2]

Circuit arrangements for several typical vacuum-tube voltmeters are illustrated in Fig. 247. The arrangement (*a*) is an ordinary anode detector in which the change in d-c plate current is used as a measure of the

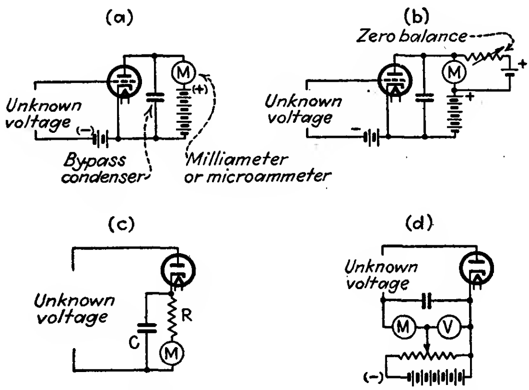


FIG. 247.—Typical vacuum-tube voltmeter circuits.

voltage applied to the grid. The circuit shown at (*b*) is a modification of (*a*) in which the residual d-c current that is present in the absence of an applied signal voltage is balanced out of the meter by the arrangement shown, thereby enabling the indicating instrument to read the increment of plate current directly. In these arrangements the relationship between the alternating voltage being measured and the resulting d-c current produced in the plate circuit depends upon the adjustment of the bias. If the tube is operated with a bias appreciably less than cut-off and the signal voltage is limited in amplitude, so that operation is always on the

[1] For further information on vacuum-tube voltmeters see the author's book, "Measurements in Radio Engineering," 1st ed., pp. 18–31.

[2] At rather high, but not extremely high, radio frequencies the vacuum-tube voltmeter will have substantially the same calibration as at low frequencies, but the transit time of the electrons will cause power to be consumed by the tube even when no grid current flows. When a plate detector biased to cut-off is employed, the resulting equivalent input resistance is approximately inversely proportional to frequency and is considerably higher than the input resistance of the same tube operated under ordinary amplifier conditions. See discussion by J. G. Chaffee, *Proc. I.R.E.*, vol. 24, p. 105, January, 1936.

curved part of the characteristic as illustrated in Fig. 246, the detector operates as a square-law device and so gives a change in d-c plate current that is exactly proportional to the square of the effective value of the applied voltage. On the other hand, if the bias approximates cut-off, then for signal voltages that are not too large the change in d-c plate current produced by the signal is very nearly proportional to the square of the effective value of the positive half cycles of the signal, giving what is termed a half-wave square-law characteristic. Finally, if the bias is appreciably greater than cut-off, the indications are determined largely by the peak amplitude of the wave. With adjustments that give half-wave square-law or peak action, reversing the polarity of the applied voltage will in general change the reading of the instrument when the applied voltage is not a sine wave. This effect is known as *turnover*, and can be avoided by operating the vacuum tube as a full-wave square-law device.

A diode detector can be used as a peak vacuum-tube voltmeter by means of the arrangements illustrated in Fig. 247*c* and 247*d*. In the circuit of Fig. 247*c* the use of a high load resistance R will make the direct-current voltage developed across this resistance only very slightly less than the crest amplitude of the applied signal, so that a microammeter M placed in series with the load resistance R will read peak voltages. A modification of this arrangement is illustrated in Fig. 247*d*. Here the signal is applied to the diode and a positive bias voltage V is applied to the cathode and adjusted until the microammeter M just shows signs of current. The peak voltage is then equal to the cathode bias V.

Detection of Frequency-modulated Waves.—Frequency-modulated signals are detected by first converting into amplitude modulation and then rectifying in the usual manner. The conversion from frequency to amplitude modulation is accomplished by passing the frequency-modulated signals through selective circuits that discriminate against one side band in favor of the other, as, for example, by detuning the receiver slightly off resonance so that the response is different for the two side bands. Receivers for frequency-modulated waves are discussed in greater detail in Sec. 113.

Electron-oscillator Detectors.[1]—Electron oscillators, either of the Barkhausen or of the magnetron type, are among the most sensitive detectors at extremely high frequencies. Typical circuit arrangements of such detectors are illustrated in Fig. 248, and are seen to be essentially ordinary electron oscillators of the type discussed in Sec. 71, with means provided by which energy from the incoming signal is superimposed upon the

[1] See Kinjiro Okabe, The Amplification and Detection of Ultra-short Electric Waves, *Proc. I.R.E.*, vol. 18, p. 1028, June, 1930; Nello Carrara, The Detection of Microwaves, *Proc. I.R.E.*, vol. 20, p. 1615, October, 1932.

oscillations. The injection of this energy into the circuits of the electron oscillator influences amplitude of oscillations and thereby affects the d-c current flowing to the anode, giving an effect equivalent to a rectification. The best results are obtained when the frequency of the electron oscillations is approximately the same as the frequency of the incoming signal, although the exact adjustment is not highly critical and marked rectifying action is present even with low-frequency signals.[1]

86. Comparison of Detection Methods.—The diode detector is by far the most widely used of all detectors because there is no practical limit to the voltage it can handle, because it has lower distortion than other detection methods when properly proportioned, and because it develops a

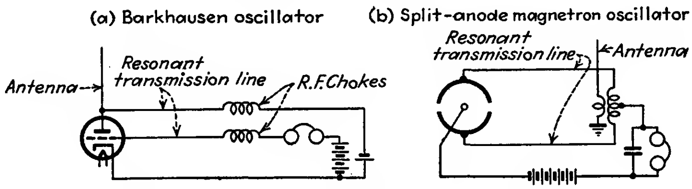


FIG. 248.—Circuits for electron-oscillator detectors.

direct-current rectified voltage of the proper polarity for use in automatic-volume-control systems. The only disadvantage of a diode detector is the fact that it absorbs appreciable power from its exciting circuit, but this is largely compensated for by the fact that the diode rectifier can be supplied by an added section built into an ordinary triode or pentode tube, and the combination of diode and tube gives an over-all sensitivity that is fully as great as obtained from any other method of detection, even taking into account the input losses.

The plate detector has the advantage over the diode in that it consumes no input power, but, because of the relatively low conversion transconductance of the tube when biased for detector operation, the over-all sensitivity is usually less than that of the same tube provided with a diode section and used to amplify the diode output. The plate detector has a limited power-handling capacity, has higher distortion than the diode, and cannot directly provide a voltage for automatic-volume-control purposes. It has therefore been largely displaced by the diode.

The grid-leak power detector is similar to the diode with respect to input resistance, distortion, and the fact that its output voltage is of the proper polarity for automatic-volume-control use, but the power-handling capacity of grid-leak power detectors employing triode and pentode tubes is so low that it is impracticable to obtain sufficient automatic-volume-control voltage without overloading the detector. The Wunder-

[1] See H. E. Hollmann, The Retarding Field Tube as a Detector for Any Carrier Frequency, *Proc. I.R.E.*, vol. 22, p. 630, May, 1934.

lich tube increases the power limit to the point where practical automatic volume control is possible, but the power-handling capacity is still limited and the diode is generally considered preferable.

Square-law detectors are characterized by high distortion when rectifying modulated waves, and have low detection efficiency. They are also of limited power capacity if the exciting voltage is restricted to a strictly square-law part of the tube characteristic. Square-law detectors are used primarily in heterodyne detection, where this method of rectification is distortionless, and in laboratory equipment, particularly vacuum-tube voltmeters.

87. Heterodyne Detection.—When two signals of slightly different frequencies are superimposed, the envelope of the resulting oscillation varies in amplitude at a frequency that is equal to the difference between the frequencies of the two alternating currents, and swings through an amplitude range equal to the crest amplitude of the smaller of the two voltages, as is shown in Fig. 249. This result is obtained because at one moment the two waves will be in phase and so will add together, while a short time later the higher frequency wave will be one-half cycle ahead of the other wave and so will combine with it in phase opposition. The rate at which the amplitude of the envelope varies is called the beat frequency (or the difference frequency), and the production of such beats by combining two waves is known as heterodyning. Since rectification of such a heterodyne signal gives a rectified current that varies in amplitude at the beat frequency, *heterodyne action gives a means of changing the frequency of an alternating current.*

The procedure for changing the frequency of an unmodulated wave by heterodyne action is to superimpose upon this signal a local oscillation having a frequency that differs from that of the signal by the desired frequency. The local oscillation may have either a higher or lower frequency than does the signal, since it is only the difference that is important. This heterodyne signal is then applied to a detector, and the desired beat frequency will be contained in the rectified output. If the wave that is to have its frequency changed is modulated, the amplitude of the beats that are produced by the superposition of the local oscillation will vary in accordance with the amplitude of the modulated wave, and the final result of the heterodyne operation is to change the frequency of the carrier wave to the new beat frequency without disturbing the character of the modulation.

The heterodyne principle of frequency changing has a number of important applications in radio communication. It can be used to change the carrier of a radio-telegraph signal to an audible frequency such as 1000 cycles, which can be used to actuate a telephone receiver. This result is accomplished by making the difference between the signal and

local oscillation frequencies a suitable audio frequency, and is known as heterodyne code reception. Another application of the heterodyne principle is in the superheterodyne type of radio-frequency amplification, in which the heterodyne principle is used to change the carrier frequency of the radio signal to a predetermined and readily amplifiable radio frequency at which the amplification takes place. In this way the

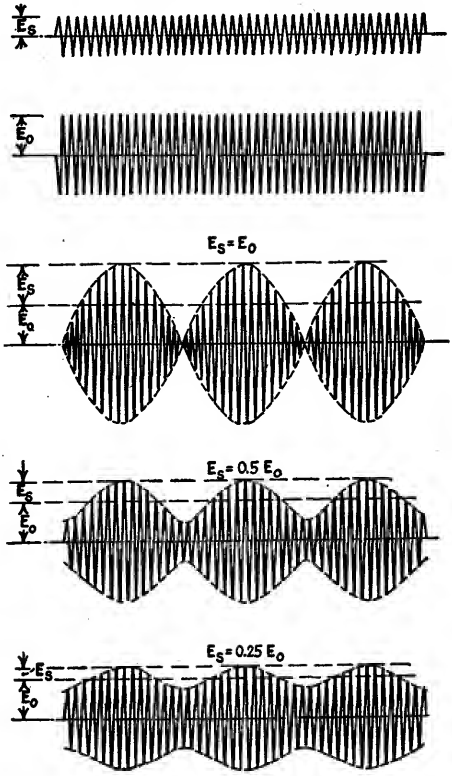


FIG. 249.—Typical heterodyne waves, showing how the combining of two waves of slightly different frequencies results in a wave which pulsates in amplitude at the difference frequency of the component waves, and how the wave shape of the envelope of the resultant wave depends upon the relative amplitudes of the two components.

frequency of the signal is changed to fit the amplifier, rather than the amplifier adjusted to fit the signal. Heterodyne action can also be employed to separate frequencies that differ from each other by a relatively small percentage. Thus it would be practically impossible to separate currents having frequencies of 1,000,000 and 1,001,000 cycles by the use of tuned circuits, but this can be readily accomplished by the use of heterodyne action. For example, if the local oscillation has a frequency of 999,000 cycles, the heterodyne action will change the original

frequencies to 1000 and 2000 cycles, respectively, which can be separated with ease.

Analysis of Heterodyne Detection.—The object of heterodyne detection is to produce sinusoidally varying currents of the difference frequency. When this result is accomplished the detection is distortionless, while, if the detector also produces harmonics of the beat frequency, distortion is introduced. The characteristics required for distortionless detection can be determined by considering the equation of the envelope of a heterodyne signal. When two sine waves of amplitudes E_o and E_s and having a difference frequency $\omega/2\pi$ are superimposed, the shape of the resultant envelope depends upon the relative amplitudes of the two waves, as shown in Fig. 249, and has the equation[1]

$$\text{Instantaneous amplitude of envelope} = \sqrt{E_s^2 + E_o^2 + 2E_sE_o \sin \omega t} \qquad (190)$$

The shape of the resulting envelope is identical with the output wave of a full-wave rectifier when the superimposed voltages are of equal amplitude, and approaches a sine-wave variation only in the limit when one of the components is extremely small compared with the other. *The shape of the envelope is such that square-law detection, that is, rectification in which the output is proportional to the square of the envelope, gives distortionless heterodyne detection, while linear rectification, by reproducing the envelope of the heterodyne signal, fails to do so.*

The character of the distortion that results from linear detection of a heterodyne signal can be obtained from a Fourier analysis of the envelope equation.[2] When this is done, it is found that the difference-frequency component of the output is largely independent of the amplitude of the stronger signal component E_o and is nearly proportional to the strength

[1] In its most general form the equation of the envelope of a heterodyne signal can be written as

$$e = E_o \sin \omega t + E_1 \sin [(\omega + \delta_1)t + \phi_1] + E_2 \sin [(\omega + \delta_2)t + \phi_2] + \cdots \qquad (191)$$

where ϕ_1, ϕ_2, etc., are phase angle constants and δ_1, δ_2, etc., are 2π times the frequency by which their respective terms differ from the frequency of the E_o term. The envelope of Eq. (191) can be found by solving Eq. (191) for zero, equating to zero the *partial derivative with respect to* ω of this transposed equation, and simultaneously solving the resultant equation with Eq. (191) to eliminate ω. Carrying out these operations gives the result

$$\begin{aligned}\text{Envelope} = \{&E_o^2 + E_1^2 + E_2^2 + \cdots \\ &+ 2E_oE_1 \cos (\delta_1 t + \phi_1) + 2E_oE_2 \cos (\delta_2 t + \phi_2) + \cdots \\ &+ 2E_1E_2 \cos [(\delta_1 - \delta_2)t + (\phi_1 - \phi_2)] + \cdots \}^{1/2}\end{aligned} \qquad (192)$$

When only two components are present and ϕ_1 is taken as zero, this equation reduces to Eq. (190).

[2] For information as to details of such analyses, see F. E. Terman, Linear Detection of Heterodyne Signals, *Electronics*, vol. 1, p. 386, November, 1930.

of the weaker component E_s. The magnitude of the deviation from this approximate relation is indicated by the fact that increasing the stronger signal component from equality with the weaker component to a value many times the weaker component, while holding the latter constant, increases the difference-frequency output by approximately 18 per cent. The amplitude of the distortion frequencies produced in linear detection of heterodyne signals is greatest when the two signal components are of equal size, under which condition the second harmonic of the beat frequency is 20 per cent.

When the signal that is to have its frequency changed by heterodyne action contains several frequency components, square-law detection of the heterodyne signal produces an output that contains every possible difference frequency that is present in the heterodyne signal, and these various difference-frequency components of the output each have an amplitude proportional to the product of the amplitudes of the two waves producing the difference frequency.

These considerations relating to the relative behavior of square-law and linear detection of heterodyne signals show that square-law detection is to be preferred from the point of view of avoiding distortion, and that if linear detection is employed the superimposed local oscillation should have an amplitude that is much larger than that of the signal which is to have its frequency changed. Under practical circumstances linear detection is usually employed because the amplitude of the local oscillation is relatively large, and it is almost impossible to maintain square-law action when this is the case.

The output of a detector that is rectifying a heterodyne signal always contains components other than the difference-frequency current and its harmonics. When the heterodyne signal contains two components, the most important of such new frequencies has a frequency that is the sum of the two frequencies being combined, while complex combination frequencies also are usually present in small amplitudes. The exact nature of these additional products of rectification depends upon the detector characteristics and is of little importance since these components have not been found to have practical usefulness.

88. Converters for Superheterodyne Receivers.—The superheterodyne receiver is so widely used that special tubes and methods have been devised for performing the frequency-changing operation involved. The heterodyne detector for such purposes, commonly referred to as first detector, converter, or mixer, is required to develop a difference frequency ordinarily in the range 75 to 500 kc by combining the incoming signal with a local oscillation differing in frequency by the desired amount. The principal arrangements employed in practice are plate detection, the 6L7 mixer tube, and the pentagrid converter.

The plate-detector type of mixer is illustrated in Fig. 250, and is an ordinary pentode plate detector biased approximately to cut-off, with provision for introducing a locally generated voltage in the tube circuits. The output load impedance is a tuned circuit resonant at the difference frequency to be produced, and is designed exactly as in any ordinary tuned amplifier. The local oscillator may be of any type, and can be coupled into the mixer tube in any convenient way, but preferably in such a manner that the voltage introduced is approximately constant as the frequency of the oscillator is varied to accommodate signals of different frequencies. For best results the amplitude of the local oscillator voltage applied to the tube should be as large as possible without overloading the detector, since in this way the conversion transconductance will be at a maximum. With pentode mixer tubes the detector plate resistance is so

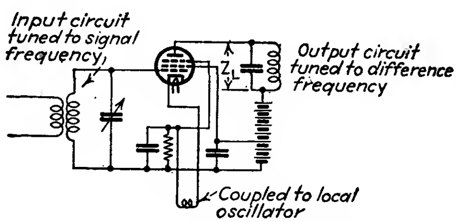


FIG. 250.—Circuit of typical plate-detector type of mixer tube.

high that it can be considered infinite, so that the difference-frequency output is given by the equation

$$\left.\begin{array}{l}\text{Output carrier voltage developed}\\ \text{across plate load}\end{array}\right\} = S_c Z_L E_0 \qquad (193)$$

where

E_0 = carrier amplitude of applied signal
Z_L = load impedance in plate circuit to difference frequency
S_c = conversion transconductance.

The conversion transconductance in this equation is the same conversion transconductance discussed in Sec. 84, and so can commonly be expected to be in the neighborhood of 0.3 to 0.4 of the mutual conductance of the same tube operated as an ordinary amplifier.

The plate-detector type of mixer has the disadvantage of introducing coupling between the local oscillator and the circuit tuned to the incoming signal. This can cause undesirable interaction, particularly in the high-frequency bands of all-wave receivers, where the local oscillator and signal frequencies differ by only a small percentage. Under such circumstances adjustment of the input tuned circuit will affect the oscillator frequency, and strong interfering signals having frequencies differing only slightly from the local oscillator frequency will tend to make the latter synchronize

automatically with the interfering signal (see Sec. 69). As a consequence the plate-detector type of mixer, while perfectly satisfactory for broadcast and lower frequencies, has very important shortcomings when used at higher radio frequencies.

The Hexode Mixer Tube (or 6L7).[1]—The hexode mixer tube or 6L7 completely isolates the local-oscillator and signal-frequency circuits by the expedient of utilizing the local oscillator to suppressor-grid modulate the amplified signal-frequency currents. The 6L7 tube contains five grids connected as shown in Fig. 251. The first or inner grid G_1 is the normal control grid, designed to have a variable-mu characteristic, and has the signal voltage applied to it. The next grid G_2 is an ordinary screen grid while the third grid G_3 is a suppressor grid that is used to suppressor-modulate the electron stream produced by the control and screen grids G_1 and G_2. The next grid G_4 is a screen grid while G_5 is a suppressor grid that is connected to the cathode. It will be noted that this arrangement is essentially a suppressor-grid-modulated amplifier, modified by the addition of grids G_4 and G_5 in order to make the plate resistance of the tube similar to that obtained with an ordinary pentode. The arrangement is capable of producing a difference frequency, since, when the local oscillator is modulated upon the incoming signal, the lower side band has a frequency that is the difference between the signal or carrier frequency and the modulating or local oscillator frequency.

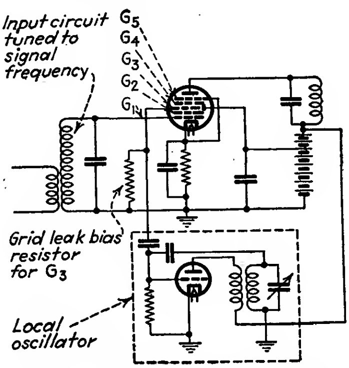


FIG. 251.—Circuit of typical 6L7 mixer tube.

The plate resistance of the 6L7 mixer is so high that it can be considered as infinite for all practical purposes, so that the difference-frequency output can be expressed in terms of the load impedance in the plate circuit and the conversion transconductance by Eq. (193). The value of the conversion transconductance of the 6L7 mixer tube depends upon the total space current of the tube and upon the negative bias and oscillator voltages applied to the grid G_3. For best results this grid should have a direct-current bias corresponding to plate current cut-off, while the oscillator voltage superimposed should be sufficient to drive this grid G_3 positive. The necessary bias is generally obtained by some self-bias arrangement such as that of Fig. 251. Under such conditions the con-

[1] For further information see C. F. Nesslage, E. W. Herold, and W. A. Harris, A New Tube for Use in Superheterodyne Frequency Conversion Systems, *Proc. I.R.E.*, vol. 24, p. 207, February, 1936.

version transconductance is of about the same numerical magnitude as in other types of converters.

The Pentagrid Converter.—The pentagrid converter is a combined oscillator and detector tube. A typical circuit arrangement is shown in Fig. 252, where the cathode, first grid G_1, and second grid G_2 function as an ordinary triode oscillator with grid G_2 acting as the anode electrode. This oscillator serves to control the flow of electrons from the cathode, causing the current to travel toward the plate in pulses occurring at the peak of each cycle of the oscillator. Most of the electrons in the pulse drawn from the cathode pass through the spaces between the wires of G_2, pass the screen-grid G_3 which is for the purpose of providing an electrostatic shield, and come to rest in front of the grid G_4, where a virtual cathode is formed. The signal voltage is applied to G_4 and so controls the number of electrons that the plate is able to draw from this virtual cathode. Grid G_5 is a screen grid that serves to make the plate current substantially independent of plate voltage, thereby giving the tube a high plate resistance corresponding to that obtained with a screen-grid tube. The virtual cathode in front of G_4 forms with each pulse of space current and then disappears between pulses. As a consequence the current actually arriving at the plate is modulated by both oscillator and signal voltages, giving a result equivalent to modulating the signal upon the oscillator frequency, thereby developing a difference frequency as one side band.

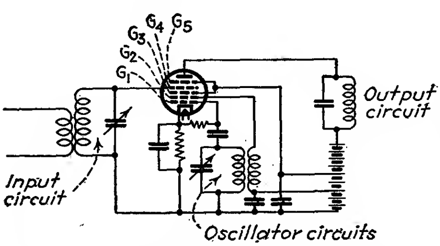


FIG. 252.—Circuit of typical pentagrid converter tube.

The screen grid G_5 makes the effective plate resistance of the detector extremely large, so that the load circuit is designed exactly as in the case of a pentode or screen-grid tube, and the output voltage can be expressed in terms of the load impedance and an equivalent transconductance according to the Eq. (193). The numerical value of the conversion transconductance under ordinary conditions is approximately that of a corresponding plate detector or 6L7 mixer. In the usual pentagrid converter, grid G_4 is provided with a variable-mu characteristic, making it practicable to control the conversion transconductance and hence the output by varying the bias on this grid.

The pentagrid converter has the merit of simplicity in that it avoids the necessity of a separate oscillator tube and also eliminates much wiring that would otherwise be required. Its usefulness is limited, however, by residual coupling between the oscillator and signal sections of the tube. This is in spite of the screen grid G_3, and is the result of capacity between

G_4 and the space charge of the virtual cathode formed in front of this grid. Inasmuch as the space charge pulsates at the oscillator frequency, the effect of this capacity coupling to the space charge is to cause oscillator-frequency currents to flow from grid G_4 through the tuned input circuit to ground. Trouble from this cause becomes more pronounced at high frequencies and as the percentage difference between local oscillator and signal frequencies is reduced.[1] The result is that the pentagrid converter, while entirely satisfactory for use with signals of broadcast and lower frequencies, becomes increasingly unsatisfactory with signals of higher frequency.

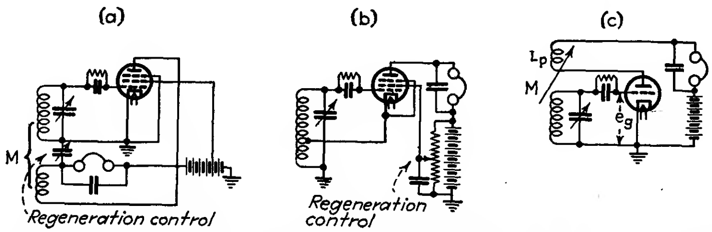


FIG. 253.—Typical circuits for regenerative and oscillating detectors. The telephone receivers indicated in the figure can be replaced by an amplifier when further amplification is desired.

89. Regenerative and Oscillating Detectors.—In examining the action taking place in detectors employing either grid or plate rectification it will be noted that there are signal currents flowing in the plate circuit of the detector, in addition to the products of rectification, as is clearly shown in Figs. 240 and 244. It is possible to obtain regeneration by feeding back a portion of this signal energy to the circuits associated with the detector input by means such as illustrated in Fig. 253. Regeneration produced in this way by utilizing the radio-frequency energy in the detector plate circuit can be more readily controlled than regeneration in amplifiers, and is occasionally used to increase the amplification and selectivity of radio receivers.

Regeneration in detectors produces exactly the same action as regeneration in radio-frequency amplifiers, since the regenerative detector is essentially a radio-frequency amplifier as far as the feedback is concerned. The effect of regeneration, no matter how produced, is equivalent to altering the effective resistance and effective reactance of the input circuit. The change of reactance caused by regeneration alters the resonant frequency slightly, making the resonance point somewhat dependent upon the adjustment of the regenerative control, but this

[1] There is also a certain amount of coupling between oscillator and signal circuits even at low frequencies as a result of the fact that the voltage on the signal grid affects the current flowing to the oscillator grid G_2. See Paul W. Klipsch, Suppression of Interlocking in First Detector Circuits, *Proc. I.R.E.*, vol. 22, p. 699, June, 1934.

effect is small because the reactance change amounts to only a few ohms and so is small compared with the large reactances in the tuned circuit. The change of resistance resulting from regeneration is much more important because the resistance of the tuned input circuit is so low that a few ohms added or subtracted represents a large percentage variation. When the energy is fed back in the proper phase to reinforce the applied signal, the effect is to neutralize a part of the resistance of the tuned input circuit. This raises the effective Q and so increases the resonant rise of voltage (which is equivalent to added amplification), as well as making the selectivity greater.

When the regeneration is carried as far as possible (*i.e.*, until the effective resistance of the input circuit approaches zero), the resulting amplification is very great for extremely weak signals, and still large but less for strong signals.[1]

While representing an inexpensive means of obtaining radio-frequency amplification, regeneration has several disadvantages. In the first place regenerative amplification is obtained by lowering the effective resistance of a tuned circuit, and, since this also greatly increases the selectivity, regeneration tends to suppress the higher side-band frequencies contained in the signal. In the second place the adjustments required to give satisfactory regenerative amplification also depend upon the frequency of the signal so that it is necessary to readjust the regeneration controls for every new signal. Furthermore the adjustments required to give appreciable regenerative action are rather critical, and a certain amount of skill is required to carry them out properly. Finally, when the regenerative action is carried to the point where the circuit resistance is completely neutralized and becomes negative, as will inevitably occur from time to time as the result of accidental improper adjustments, oscillations will be set up which will heterodyne with any signal that may be present and produce annoying squeals. These disadvantages of regenerative amplification are so great that it is generally considered better practice to obtain radio-frequency amplification by the use of tuned radio-frequency amplifiers rather than by regeneration.

Oscillating Detectors.—When regeneration is increased to the point where the resistance of the resonant circuit is completely neutralized,

[1] This dependence upon signal voltage arises because the third-order curvature of the tube characteristic causes the effective plate resistance of the tube to be slightly greater for large signals than for small, thus reducing the regeneration for large signals. Analysis shows that, when the regeneration is made as great as possible without oscillation in the absence of a signal, the amplification is inversely proportional to the two-thirds power of the signal voltage and directly proportional to the response obtained when no regeneration is present. See Balth. van der Pol, The Effect of Regeneration on the Received Signal Strength, *Proc. I.R.E.*, vol. 17, p. 339, February, 1929.

there will be set up oscillations which will heterodyne with any signal currents in the resonant circuit. The resultant beats are rectified by the detector and cause difference-frequency currents to appear in the detector output. The oscillating detector therefore acts as a heterodyne detector in which the detector tube generates the heterodyne oscillations, as well as functioning as a rectifier. The circuits used in oscillating detectors are the same as those employed for regenerative detectors, the only difference being that the regeneration is increased to the point where oscillations are produced. Grid-leak power rectification is always employed in oscillating detectors because of its high sensitivity and because it automatically supplies the proper grid bias for the oscillations.

The oscillating detector is used extensively in the reception of code signals, particularly short-wave code signals. Compared with a separate heterodyne for such purposes, the arrangement has the advantage of much greater sensitivity[1] and of being simpler to adjust. Thus, when the local oscillations have a frequency suitable for heterodyning the code signal to an audio-frequency such as 1000 cycles, the input circuit is automatically tuned approximately to resonance with the incoming signal. The oscillating detector arrangement is not suitable for use with superheterodyne receivers, however, and is also usually avoided in code reception of the lower radio frequencies. This is because under such conditions the difference frequency is a large percentage of the signal frequency, so that when the oscillations have the proper frequency the input circuit is very considerably detuned from the signal frequency.

The great sensitivity of the oscillating detector is a result of the large regenerative amplification that the signal undergoes before being rectified. When no signal is present, the oscillations have an amplitude such that the effective plate resistance of the tube has a value for which the regeneration exactly neutralizes the resistance of the resonant circuit. When this condition exists, the regenerative amplification to a superimposed oscillation is very great because the situation is much the same as that which exists in an ordinary regenerative detector adjusted to give the maximum possible regeneration. The difference in the two cases, however, is that with the regenerative detector this adjustment is very critical and impossible to maintain, whereas in the oscillating detector the oscillations automatically assume an amplitude that picks out this critical condition and maintains it with complete stability.

The regenerative amplification which the signal undergoes in an oscillating detector can be analyzed by considering that the signal repre-

[1] Thus a simple oscillating detector will make practically any short-wave code signal that is above the noise level audible in a telephone receiver. Effective amplifications as high as 15,000 are indicated under optimum conditions. See H. A. Robinson, Regenerative Detectors, *QST*, vol. 17, p. 26, February, 1933.

sents a voltage that is induced in the resonant circuit in addition to the voltage induced by the feedback from the plate circuit. The phase of the signal with respect to the oscillation changes from aiding to opposition at a rate corresponding to the difference frequency, as in the case of any heterodyne signal. To a first-order approximation the variation in the amplitude of the resultant wave applied to the grid of the tube acts as though there were no signal voltage applied to the detector, but rather as though the regeneration of the oscillating detector were alternately decreased and increased from its actual value at a rate corresponding to the beat frequency. The oscillating detector hence has its greatest sensitivity when a small change in the regeneration will produce a large change in the amplitude of the generated oscillations. This condition is always realized when the regeneration has the smallest value at which oscillations will exist and when the resonant circuit has the highest possible Q (*i.e.*, lowest possible actual resistance).

Typical oscillating-detector circuit arrangements are illustrated in Fig. 253. Pentode tubes are generally used, with the regeneration control obtained by varying either the screen voltage or the electrostatic or magnetic coupling between input and plate circuits. When the regeneration control is applied to the screen voltage, the circuit design should be such that oscillations will stop when the screen is of the order of 20 to 40 volts. If the stopping point is at higher potentials, the oscillations start and stop with an annoying thump, whereas, if the critical point is a lower voltage, the conversion efficiency of the arrangement as a detector is low because of the small plate current.

When the oscillating detector employs a triode tube, trouble is sometimes encountered from a sustained audio-frequency sound which occurs when the adjustment is such that oscillations are just barely maintained. This is known as "threshhold" or "fringe" howl, and is to be avoided since it occurs under conditions for which the oscillating detector is most sensitive. Threshhold howl may occur in grid-leak arrangements when the audio-frequency load impedance in the plate circuit of the triode is inductive, and it can be cured either by using resistance coupling or by shunting the inductive load with a sufficiently low resistance. Threshhold howl does not ordinarily occur when pentode tubes are employed.[1]

90. Superregenerative Detectors.[2]—A superregenerative detector is a regenerative detector which is varied from an oscillatory to a non-oscil-

[1] The mechanism of threshhold howl is described by L. S. B. Alder, Threshhold Howl in Reaction Receivers, *Exp. Wireless and Wireless Eng.*, vol. 7, p. 197, April, 1930.

[2] For further information on superregeneration, see Hikosaburo Ataka, On Superregeneration of an Ultra-short Wave Receiver, *Proc. I.R.E.*, vol. 23, p. 841, August, 1935; D. Grimes and W. S. Barden, A Study of Superregeneration, *Electronics*, p. 42,

lating condition at a low radio-frequency rate. During the oscillatory interval, oscillations build up, only to be suppressed, or "quenched," and the resulting action is such that with proper adjustment an applied signal is amplified enormously before detection. A typical superregenerative circuit is shown in Fig. 254, and consists of a tube arranged to regenerate in the manner shown in Fig. 253*c*, but supplied with a plate voltage that is a low radio frequency, such as 25 kc. Oscillations then build up during the half cycles when the plate is positive, but die out (*i.e.*, are quenched) during the time the plate is negative. For proper operation the oscillations must die out completely before they start to built up, which is equivalent to saying that the average resistance of the circuit must be positive.

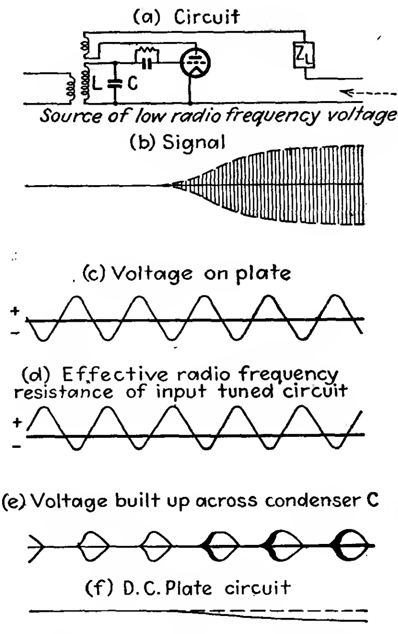


FIG. 254.—Simple superregenerative circuit, together with details showing the mechanism by which superregenerative amplification is obtained.

When no signal is present, the initial pulse that starts the building up of oscillations is supplied by thermal agitation, shot effect, etc., and the resulting oscillations are as illustrated to the left in Fig. 254*e*. The area under the envelope of the curve of oscillations depends upon the amplitude of the initiating pulse, and, since this is a chance factor of thermal agitation, etc., the areas under successive envelopes differ in a random manner, and the rectified output contains a characteristic hiss. However, upon the application of a signal that has a greater amplitude than the random voltages, the signal then becomes the initiating pulse, thereby suppressing the characteristic hiss and causing the building-up process to get under way faster, as shown to the right in Fig. 254*e*, where the shaded area represents the difference caused by the presence of the signal. Inasmuch as the initiating pulse, and hence the time required to reach full amplitude, is proportional to the amplitude of the signal, the shaded area in Fig. 254*e* will vary with the amplitude of the signal, and hence will reproduce in a rough sort of way the modulation of the signal.

February, 1934; Edwin H. Armstrong, Some Recent Developments of Regenerative Circuits, *Proc. I.R.E.*, vol. 10, p. 244, August, 1922; M. G. Scroggie, The Superregenerative Receiver, *Wireless Eng.*, vol. 13, p. 581, November, 1936.

The output of the superregenerative detector is obtained by rectifying the oscillations that are built up across the tuned circuit, using a grid-leak condenser combination as shown in Fig. 254*a*. The bias voltage developed across this combination will be proportional to the average amplitude of the oscillations, and so will vary in accordance with the shaded area as illustrated in Fig. 254*f*, causing the plate current to be reduced by the presence of a signal by an amount that varies in accordance with the modulation envelope.

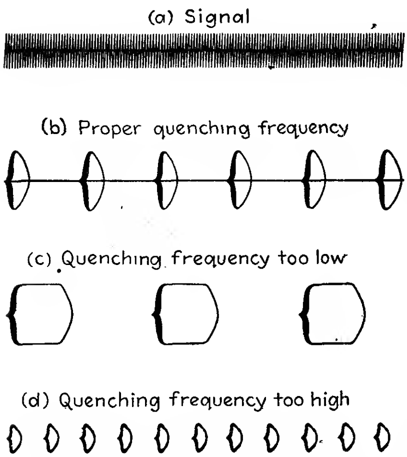


FIG. 255.—Oscillograms showing effect of quench frequency upon output of superregenerative detector. The shaded area represents the output caused by the presence of a signal.

In order to obtain best results with a superregenerative detector it is necessary to have a proper balance between the circuit proportions and the amplitude and frequency of the quenching oscillation. The effect of the quenching frequency is illustrated in Fig. 255. Starting with low quenching frequencies, it is seen that at first increasing the quenching frequency increases the number of times the oscillations build up in a given length of time and therefore increases the total of all the shaded areas almost in direct proportion to the quenching frequency. The output and hence the sensitivity therefore increase with quenching frequency, until an optimum value is reached such that the oscillations just have time to approach full amplitude when they are quenched. With quenching

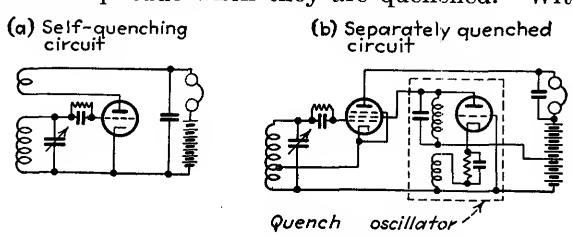


FIG. 256.—Circuits of typical superregenerative detectors.

frequencies higher than this, the width of the shaded area decreases and there is a tendency for the oscillations to be quenched before they have time to build up to full amplitude, so that, although the number of shaded areas per second increases, the size of each area is so diminished that the net result is a considerable loss in sensitivity.

Two additional circuit arrangements for producing superregeneration are illustrated in Fig. 256. The arrangement at (*a*) is particularly

useful because it requires no special circuits or auxiliary quenching oscillator. This circuit obtains the quenching action by operating the circuit so that interrupted oscillations are produced by the mechanism discussed in Sec. 66, thereby making the arrangement self-quenching. The interrupted oscillations can be obtained by use of a high grid-leak resistance, a relatively large grid-condenser capacity, and sufficient regeneration to make the circuit strongly oscillatory. The frequency of interruption can be controlled by varying either the regeneration or the grid-leak resistance.

A properly adjusted superregenerative detector is characterized by extremely great sensitivity. A single tube is capable of giving an audible output with signals that have an amplitude comparable with the thermal agitation voltages present in the input circuit. Experience with superregenerative detectors also indicates that they are much less susceptible to such interference as ignition noises than are most receivers. The reason for this appears to be that there is an inherent limiting action contained in the mechanism of operation whereby very loud signals produce only slightly more output than do weak signals. As a consequence strong intermittent noise voltages produce outputs only slightly larger than those obtained from weak signals.

At the same time the superregenerative detector has a number of limitations. In the first place, a characteristic hiss is always present in the absence of an applied signal, and, though this hiss disappears in the presence of a signal, it makes superregeneration impracticable for such purposes as broadcast receivers. A superregenerative receiver also possesses rather poor selectivity, this being necessarily true because of the large number of side-band frequency components produced by the quenching action. The principal practical use of superregeneration has been in the reception of signals of such high frequency that ordinary methods of amplification cannot be employed.

91. Detector Output When the Applied Signal Consists of Two Modulated Waves.—Circumstances commonly arise where in addition to the desired signal there is also a relatively weak but not negligible interfering signal applied to the detector input. The most important effects produced under such conditions are: (1) the suppression of the weaker signal when a linear detector is employed and when the difference in carrier frequencies of the two signals is above audibility; (2) an audible beat note representing the difference frequency between the two carriers when these differ in frequency by an amount lying in the audible range; (3) a flutter occurring when the carrier frequencies are almost but not exactly in synchronism; and (4) distortion occurring when the carriers are of identical frequency and when both desired and undesired stations are broadcasting the same program.

When two modulated signals of unequal amplitudes and of carrier frequencies that are so different as to produce an inaudible beat frequency are simultaneously applied to the input of a linear rectifier, it is found that the weaker of the two signals is not rectified. This is because under these conditions the envelope of the combined signal is as shown in Fig. 257, having a principal modulation representing the desired modulation and a minor variation corresponding to the inaudible beat frequency between the two carriers. This minor variation is modulated in accordance with the modulation of the undesired carrier, but upon *linear* rectification the detector output follows the modulation envelope and hence contains no component that follows the modulation of the undesired weaker signal. *This suppression of the weaker modulation is equivalent to an increase in the effective selectivity and represents an important property of a linear detector.*[1] In order that the suppression of the weaker signal may be complete, it is necessary that the strong signal be considerably larger than the weaker and that the detector be exactly linear. With ordinary rectifiers the suppression becomes very pronounced when the ratio of signal amplitudes exceeds 2 to 1. This suppression of the weaker signal does not occur in the case of square-law detectors.

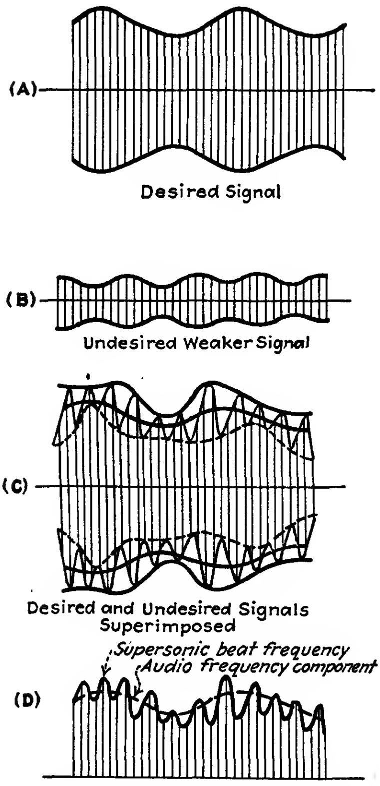


FIG. 257.—Wave forms obtained in the linear detection of a signal consisting of a weak modulated wave superimposed upon a strong modulated wave. The rectified output of the linear detector is seen to contain no component varying at the modulation frequency of the weaker signal.

When the difference between the carrier frequencies of the desired and undesired signals is in the audible range, a number of undesired

[1] For further information and details of the method of analyzing this phenomenon, see R. T. Beatty, Apparent Demodulation of a Weak Station by a Stronger One, *Exp. Wireless and Wireless Eng.*, vol. 5, p. 300, June, 1928; S. Butterworth, Note on the Apparent Demodulation of a Weak Station by a Stronger One, *Exp. Wireless and Wireless Eng.*, vol. 6, p. 619, November, 1929; E. V. Appleton and D. Boohariwalla, The Mutual Interference of Wireless Signals in Simultaneous Detection, *Exp. Wireless and Wireless Eng.*, vol. 9, p. 136, March, 1932.

components of audible frequency appear in the output of both square-law and linear detectors. The component having the largest amplitude is the difference frequency between the two carriers, and this is the most disturbing component when it exceeds 100 to 200 cycles.

When the frequencies of the two carriers differ by less than about 50 cycles, the most disturbing components in the output are the difference frequencies formed by the carrier of the strong or desired signal heterodyning with the side-band frequencies of the weaker or undesired signal to produce what can be termed *side-band noise.* When the carrier frequencies differ by only a few cycles a second, this side-band noise gives rise to the characteristic flutter commonly heard when two or more broadcast stations are simultaneously transmitting on approximately the same frequency. When the two carriers are both weak enough so that there is a background of noise, the noise level in the detector output will also flutter at a frequency corresponding to the difference frequency between the two carriers, and may in some cases produce an effect more annoying than the side-band noise.

In the event that the two signals come from stations which have their carriers synchronized and which are modulated with identical programs, the detector output will not ordinarily represent a distortionless reproduction of the original modulation unless one of the carrier amplitudes is much weaker than the other. This is because the relative phase with which the two carriers and their respective side-band components combine in the detector input depends upon the distance to the transmitter, time differences in the transmission of the program, and the side-band frequency involved. The result is that, when the carrier amplitudes are of approximately the same order of magnitude, certain side-band frequencies will tend to cancel while others will be reinforced; furthermore at certain locations the two carriers will also tend to cancel and thereby distort the envelope of the wave applied to the detector input. In order for distortion of this sort to be imperceptible under the worst practical conditions, it is necessary that one carrier have an amplitude at least four times that of the other carrier.

The exact nature of these effects which occur when the two carrier frequencies differ by only a small amount, or are synchronized, depends upon whether a linear or square-law rectifier is used, upon the relative amplitudes of the two carriers, and upon their degrees of modulation. The exact analysis is too involved to be presented here, but is to be found in the literature.[1]

[1] The most extensive work on this subject is that of C. B. Aiken. The method of analysis that he has employed in attacking the problem, as well as the essential results, are contained in the following papers: Charles B. Aiken, Theory of the Detection of Two Modulated Waves by a Linear Rectifier, *Proc. I.R.E.*, vol. 21, p. 601, April

Problems

1. Explain the detailed mechanism accounting for the fact that the efficiency of detection of a diode detector is increased by: (*a*) increasing the load resistance R, (*b*) decreasing the plate resistance of the diode.

2. Explain the detailed mechanism whereby the input resistance of a diode detector of good efficiency increases with load resistance R but is nearly independent of the diode plate resistance.

3. In a diode detector as shown in Fig. 235*a*, $\eta = 0.90$, $C = 100\ \mu\mu\text{f}$, $R = 200{,}000$ ohms,

a. Calculate and plot as a function of modulation frequency, the maximum degree of modulation that the "signal input" to the detector may have for distortionless rectification.

b. Assume that the signal voltage is induced in a tuned circuit resonant at 600 kc, having $Q = 60$ and a resonant impedance of 50,000 ohms. (1) Calculate and plot, as a function of frequency up to 20 kc off resonance, the magnitude and phase of the output impedance of the tuned circuit (*i.e.*, the impedance that the diode sees when looking toward the tuned circuit). (2) Similarly calculate and plot the relative side-band amplitude of the voltage developed across the tuned-circuit output when the diode load is removed. (3) Calculate the magnitude and phase of the load impedance which the diode offers to the tuned circuit as a function of side-band frequency up to 20 kc off resonance. (4) From (1) and (3) calculate the ratio of degree of modulation appearing at the input terminals of the diode to the degree of modulation of the voltage actually induced in the resonant circuit, up to modulation frequencies of 20 kc. (5) Calculate the maximum degree of modulation that the diode can handle without distortion, and from this determine and plot the maximum allowable degree of modulation of original induced signal that can be rectified without distortion up to modulation frequencies of 20 kc.

4. In a diode detector the input circuit has a resonant impedance of 160,000 ohms. Assuming an efficiency of detection of 0.90, calculate and plot maximum permissible degree of modulation of original induced signal without negative peak clipping as a function of the ratio of a-c to d-c impedances of the diode load for ratios between 1 and 0.5 and for d-c load resistances of 150,000, 300,000, and 600,000 ohms, and also calculate and plot the approximate r.m.s. distortion if the original induced signal is completely modulated.

5. *a.* In the diode circuit of Fig. 239, go through the calculations giving the maximum allowable degree of modulation of the original induced signal without negative peak clipping, and the calculations determining the distortion when the original signal is completely modulated.

b. On the assumption that the carrier input voltage in Fig. 239 is 20 volts crest value and is completely modulated, calculate (1) direct-current voltage developed for automatic-volume-control purposes, (2) modulation-frequency voltage developed across R_4 at moderate modulation frequencies, (3) radio-frequency voltage appearing across R_4 when the radio-frequency voltage across C_1 is 1 volt.

6. Design a plate detector using a 56 tube with $E_b = 250$ volts and having a transformer-coupled output. In this design specify grid bias and maximum allowable

---

1933; A Study of Reception from Synchronized Broadcast Stations, *Proc. I.R.E.*, vol. 21, p. 1265, September, 1933; The Effect of Background Noise in Shared Channel Broadcast, *Bell Sys. Tech. Jour.*, vol. 13, p. 333, July, 1934. Also see Hans Roder, Superposition of Two Modulated Radio Frequencies, *Proc. I.R.E.*, vol. 20, p. 1962, December, 1932.

carrier amplitude that can be handled, while allowing for complete modulation; estimate detector plate resistance and from this specify proper primary inductance of output transformer for 70.7 per cent response at 60 cycles; and estimate maximum audio output voltage obtainable with reasonable detector efficiency and a transformer turn ratio of 2.

7. In a 6C6 pentode tube with $E_{sg} = 100$ volts, the control-grid bias for cutoff is -6 volts. When a carrier of 3 volts crest is applied, the d-c plate current flows in half-cycle pulses that have a maximum value of 2 ma and an average (or direct-current) value of approximately 0.5 ma. From this information, and with reasonable assumptions as to conversion transconductance, design a resistance-coupled plate detector using a 6C6 tube with $E_{sg} = 100$ volts and $E_b = 250$ volts, calculate the amplification in the middle range of frequencies, and compare with the gain of resistance-coupled amplifier using the same tube and plate-supply voltage as given in Table VII, Chap. V.

8. *a.* Design a grid-leak power detector of suitable proportions for handling broadcast signals, using a 56 tube with transformer coupling in the plate circuit. In this design estimate the proper plate voltage from data given in tube manuals.

b. Estimate the largest completely modulated carrier that can be handled without overloading, and also the output voltage, assuming a transformer step-up ratio of 3.

9. Using the analysis of Sec. 55, derive formulas for the rectified d-c current, the modulation-frequency current, and the percentage of second-harmonic distortion of the modulation frequency which results when a sinusoidally modulated wave is applied to the grid of a pentode tube operated where the tube has a true square-law characteristic.

10. A triode vacuum-tube voltmeter is adjusted to operate over the square-law part of the tube characteristic. Demonstrate by mathematical analysis that, when the applied voltage has components at several frequencies, the rectified d-c current is proportional to the square of the effective value of the applied voltage wave and is not affected by the relative phases of the various components.

11. In the vacuum-tube voltmeter circuits of Fig. 247*a* and *b*, what would be the result of (*a*) omitting the by-pass condenser in the plate circuit, (*b*) making the condenser of insufficient capacity so that it is only a partial by-pass.

12. Two sine waves having equations $e_1 = E_1 \sin (\omega_1 t + \Phi_1)$, and $e_2 = E_2 \sin (\omega_2 t + \Phi_2)$ are applied to a square-law heterodyne detector. By the method of analysis given in Sec. 55 determine the various components of current that appear in the detector output.

13. A modulated wave having the equation $e = E_0(1 + m \cos vt) \sin \omega_0 t$ is combined with an oscillation $e_1 = E_1 \sin \omega_1 t$, and the resulting combination is applied to a square-law detector. By means of the method of analysis given in Sec. 55 determine the various current components that appear in the detector output, and show that these include a carrier of frequency $(\omega_0 - \omega_1)/2\pi$ modulated in the same manner as the original input wave.

14. In a pentagrid converter, the load impedance in the plate circuit consists of a band-pass filter such as described in Prob. 17 of Chap. V. If the conversion transconductance at the operating point is 350 micromhos, calculate and plot the output voltage as a function of the difference frequency produced, when the signal is 1 mv.

15. Tubes with variable-mu characteristics are commonly employed in converters of the plate-detection type, but are seldom used as ordinary plate power detectors of modulated waves. Explain the reasons for this.

16. In the circuit of Fig. 253*c*, assume that a signal voltage e_s is acting in series with the tuned circuit. By setting up an equivalent plate circuit for radio frequencies in which it is assumed $\omega L << R_p$, derive an expression for the voltage e_g across the tuned

circuit, taking into account the voltage that the coil L_p induces in series with the signal e_s and assuming that the resonant circuit is tuned to the frequency of e_s. From this expression write an equation giving the equivalent resistance that the circuit has in the presence of regeneration, by comparing the resonant rise of voltage actually obtained with the rise in the same circuit when regeneration is absent. Finally, discuss the effect of M on the equivalent resistance of the circuit, keeping in mind that M may be either positive or negative.

17. In an oscillating detector to which there is applied a reasonably strong signal, it is found that, as the resonant frequency of the oscillator tuned circuit is varied about the frequency of the signal, the pitch of the beat note first decreases until a low value such as 400 cycles is reached, then ceases until the oscillator circuit is tuned slightly to the other side of resonance, whereupon the low pitch reappears and becomes higher as the adjustment process continues. Explain the silence in the region where the oscillator tuned circuit is approximately in resonance with the signal.

18. In a superregenerative receiver that is to be tuned over a considerable frequency range, it will be found that the optimum value of quenching frequency will vary somewhat with the frequency to which the receiver is tuned. Discuss the factors that contribute to this.

19. In a superregenerative detector it is found that, when two signals are simultaneously applied to the detector, the stronger signal almost completely suppresses the weaker signal as far as the rectified output is concerned. Explain how the mechanism of the superregenerative detector leads to this behavior.

20. Two modulated waves having equations $e_1 = E_1(1 + m_1 \cos v_1 t) \sin \omega_1 t$ and $e_2 = E_2(1 + m_2 \cos v_2 t) \sin \omega_2 t$ are applied to a square-law detector. If the difference $(\omega_2 - \omega_1)/2\pi$ in carrier frequencies is a moderate audio frequency, enumerate the various audio-frequency components of the rectifier output current and give the relative amplitudes of each.

CHAPTER XI

SOURCES OF POWER FOR OPERATING VACUUM TUBES

92. Cathode Heating Power.—The most frequently used sources of power for heating the cathodes of tubes are commercial lighting circuits (both alternating and direct), direct-current generators, storage batteries, and dry cells and other types of primary batteries. Commercial alternating-current power is used wherever possible because of its economy and simplicity. Direct-current generators are sometimes used in radiotelephone transmitters to eliminate the possibility of hum when alternating current is used to heat the filaments. Storage batteries are employed in automobile and airplane radios. Primary batteries, such as dry cells and air-cell batteries, are used in portable equipment and where commercial power sources are not available.

When 60-cycle alternating current is used to heat the cathode of a tube, there is always the possibility of introducing 60-cycle and 120-cycle components in the plate current, and also of modulating at these frequencies any signal voltages being amplified. These effects give rise to what is commonly termed *alternating-current hum* since the result is a low-pitch hum appearing in the loud-speaker. Alternating-current hum is worst with filament-type tubes, but may also be present to some extent even when indirectly heated cathodes are used.

Alternating-current Hum in Heater-type Tubes.[1]—When the heater is operated with alternating current, hum may be introduced by alternating electrostatic and magnetic fields rising from the heater current, or as a result of capacitive coupling and leakage between heater and the grid and plate electrodes. Electrostatic fields from unshielded portions of the heater influence the plate current in the same way as does a signal voltage on the grid, and they may introduce both fundamental and second-harmonic hum. The fundamental hum depends upon the potential of the heater with respect to the cathode, and may be reduced to a low value by the use of a suitable heater bias. The second-harmonic hum, however, is not affected by this procedure, and the only sure means of eliminating hum of this sort is to shield the heater and its lead wires adequately. The magnetic field produced by the heater current will introduce second-

[1] More detailed discussion of this subject is given by J. O. McNally, Analysis in Reduction of Output Disturbances Resulting from the Alternating-current Operation of the Heaters of Indirectly Heated Cathode Triodes, *Proc. I.R.E.*, vol. 20, p. 1263, August, 1932.

harmonic hum by the same mechanism that is discussed below in connection with magnetic fields in filament-type tubes. This effect can be minimized by using a heater geometry such as a double spiral or narrow loop, which creates a minimum of external magnetic field, and also by using a high-voltage low-current heater. Leakage and capacitive couplings to the grid and plate electrodes cause current from the heater circuit to flow through the external grid-circuit and plate-circuit impedances, and thereby introduce hum voltages in these circuits. Capacitive couplings can be eliminated by the proper arrangement or shielding of heater lead wires; leakage when present is usually the result of a film of "getter" material on the glass insulation between lead wires, and it can be prevented by arranging matters so that the getter cannot be deposited where the heater wires pass through the envelope of the tube.

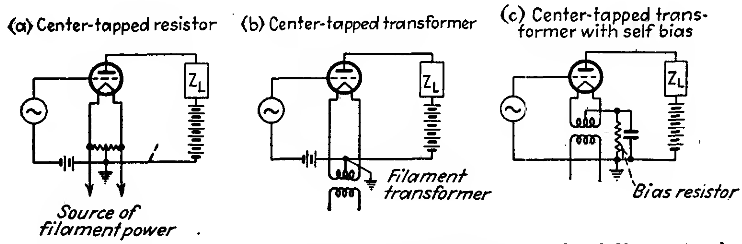


FIG. 258.—Methods of connecting the grid and plate return leads of filament tubes when alternating filament current is used.

When special care is taken, the hum in heater-type tubes can be reduced to the point where it is less than the normal thermal-agitation noises, and low hum tubes of this character are available to be used where needed. The general run of heater tubes such as are commonly used in radio receivers have somewhat more hum. The exact level depends greatly upon the design of the tube and in some cases is great enough to produce a hum voltage of a few millivolts across a high resistance in the plate circuit.

Alternating-current Hum in Filament-type Tubes.—The use of alternating current for heating the cathodes of filament-type tubes always introduces a certain amount of hum. In order to keep this within reasonable limits, it is absolutely essential that the grid and plate return leads be brought to a point that is at substantially the same potential as that of the mid-point of the filament; this may be done by one of the methods indicated in Fig. 258. Otherwise the return point would become alternately more and less positive than the average of the filament as the heating current goes through its cycle, and this is equivalent to applying voltages having the frequency of the heater current to both grid and plate circuits. As a result of the center-tapped arrangement required with alternating heating power, the bias required is $E_f/\sqrt{2}$ volts more with

a-c heating current than with direct current, where E_f is the effective value of the alternating-current voltage employed.

After the fundamental frequency component of the alternating-current hum has been eliminated by one of the arrangements shown in Fig. 258, there remains a residual hum that has a frequency twice that of the filament current. This double-frequency hum can arise from the cyclical variation of filament temperature, the effect that the alternating magnetic flux set up by the filament current has on the space current of the tube, and the effect that the voltage drop in the filament has on the space current.[1] Since the heat generated in the filament at any instant is proportional to the square of the instantaneous filament current, the filament temperature will pulsate at twice the frequency of the filament current. The heat capacity of filaments used in vacuum tubes is so high, however, that the resulting variation in filament temperature is very small with 60-cycle filament current and produces negligible hum when temperature saturation is present. The magnetic field produced by the filament current deflects the electrons flowing to the anode according to the principles discussed in Sec. 24 and causes the plate current to be slightly larger when the filament current is zero than when the current is at either a positive or a negative maximum. The voltage drop in the filament causes the negative half of the filament to supply more electrons to the anode than does the positive half, and, since the number of electrons drawn from the filament is proportional to the three-halves power of the electrostatic field, the current from the negative end of the filament is increased more by the filament drop than the current drawn from the positive end is decreased. The total space current of the tube is hence slightly greater when there is a voltage drop in the filament than when the entire filament is at the same potential.

The hum introduced by the alternating voltage drop in the filament can be reduced by using a low filament voltage and by arranging the filament in a V or W construction. Such arrangements cause electrons to be interchanged between the terminals of the filament and this action tends to smooth out the hum-producing action. The hum introduced by the magnetic field is 180° out of phase with the hum arising from the voltage drop in the filament, since the magnetic flux tends to reduce the plate current when the filament current is large while the voltage drop that accompanies this large current tends to produce an increased anode current. In filament-type tubes designed to be operated with alternating current, the hum is kept low by the use of a low filament voltage and a

[1] A more detailed analysis of the factors producing the double frequency hum is given by W. J. Kimmell, The Cause and Prevention of Hum in Receiving Tubes Employing Alternating Current Direct on the Filament, *Proc I.R.E.*, vol. 16, p. 1089, August, 1928.

filament geometry such that with normal operating conditions the magnetic and voltage-drop effects tend to balance each other.

The use of alternating current to heat filament cathodes is permissible only where the tubes involved normally carry large signal currents, as, for example, in the power stage of a radio receiver or public-address amplifier, in radio transmitters, etc. Under such conditions the desired signal is so large as to make the hum negligible in comparison. Even then, if extremely low hum level is important, as in the case of radiotelephone transmitters, it is sometimes found desirable to use direct current for heating filaments. Filament tubes operated from alternating

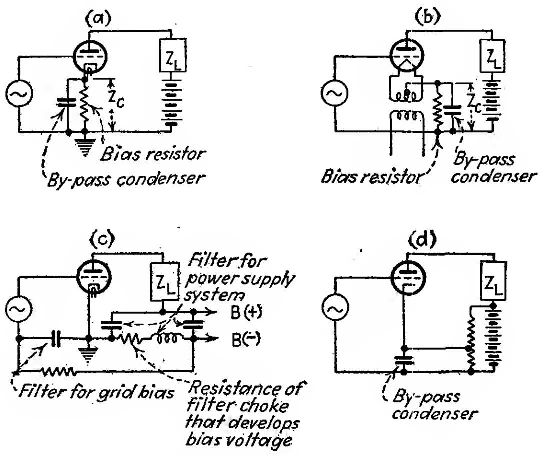


FIG. 259.—Methods of using the plate-supply voltage to make the grid negative with respect to the cathode.

current are never used in low-level audio-frequency amplifiers, and are not considered very satisfactory in low-level radio-frequency amplifiers.

93. The Grid-bias Voltage.—The grid-bias voltage for voltage amplifiers and small-power tubes is practically always derived from the plate-supply voltage by one of the arrangements illustrated in Fig. 259. The systems shown in Figs. 259*a* and 259*b*, involving a resistance between the cathode and ground in order to make the cathode positive with respect to the ground, is most commonly used, but the arrangement in Fig. 259*c* is also employed to some extent in biasing power amplifiers, since by placing the filter choke in the negative power lead, as shown, the resistance of the choke can be utilized to produce the bias.

Batteries were once universally employed for developing bias voltages, but are now used only in battery-operated equipment and in special circuits where self-bias arrangements are very complicated. Since a negative grid draws no current, the principal requirement of the bias

battery, often called C battery,[1] is a long shelf life. Bias batteries usually consist of a number of small dry cells assembled in a moistureproof carton, although special bias cells having an unusually long life under open-circuit conditions are coming into use.

Bias voltages for large-power tubes can be obtained from self-bias arrangements, direct-current generators, rectifier-filter systems, and in the case of Class C amplifiers and oscillators by grid-leak grid-condenser combinations. When direct-current generators are employed, a filter must be placed between the generator and the tube as shown in Fig. 260 in order to eliminate commutator ripple. When rectifier-filter systems are employed, it is essential that a load resistance be connected across the output of the filter, as otherwise there is no return path by which grid current can reach the cathode.

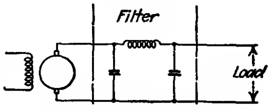


FIG. 260.—Filter circuit for eliminating ripple in output of direct-current generator.

Regeneration in Self-biased Tubes.—The self-bias arrangements of Figs. 259*a* and 259*b* have a tendency to produce regeneration because the amplified signal currents flowing in the plate circuit produce a voltage drop across the bias resistance, and this drop is applied to the grids of the tubes. In order to minimize this effect, the bias resistance is always shunted by a condenser, as shown in Figs. 259*a* and 259*b*, in order to short-circuit the bias resistance to alternating currents and prevent an alternating-current voltage from developing across it. This expedient is very effective at the higher audio and at radio frequencies where the capacitive reactance of the condenser is very small, but fails at low audio frequencies where the capacitive reactance is large.

Regeneration resulting from a grid-bias resistance acts exactly as does regeneration produced by a common plate impedance (see Sec. 51) and can be analyzed in the same way. Considering Figs. 259*a* and 259*b*, it is readily demonstrated that[2]

$$\left.\begin{array}{l}\text{Actual amplification taking into}\\ \text{account regeneration from self bias}\end{array}\right\} = \frac{A}{1 + \dfrac{\mu Z_c}{R_p + Z_L + Z_c}} \tag{194}$$

[1] The terms A, B, and C to indicate filament, plate, and grid-bias batteries, respectively, originated when vacuum tubes first began to be used. The first tube circuits required filament and plate batteries, and, in order to differentiate clearly between the two batteries, the early instructions designated the filament battery as Battery A (*i.e.*, Battery 1) and the plate battery as Battery B (*i.e.*, Battery 2), and these letters ultimately came to be used as the name of the battery. Later, when the importance of the grid-bias battery was discovered, it was naturally called a C battery (*i.e.*, Battery 3).

[2] The derivation of Eq. (194) follows: The voltage e_g acting on the grid of the tube is $(e_s + e_r)$, where e_s is the applied signal and e_r is the voltage developed across the bias impedance Z_c. The voltage e_r can be expressed in terms of e_g by making

where

A = vector amplification obtained without regeneration
μ = amplification factor of tube
R_p = plate resistance of tube
Z_c = impedance between terminals of bias resistance (impedance of combination formed by bias resistance and shunting capacity)
Z_L = load impedance in plate circuit of tube.

In pentode and screen-grid tubes when the plate resistance can be considered as substantially infinite in comparison with the load resistance, Eq. (194) reduces to

$$\left.\begin{array}{l}\text{Actual amplification taking into}\\ \text{account regeneration from self-bias}\end{array}\right\} = \frac{A}{1 + G_m Z_c} \tag{195}$$

where G_m is the mutual conductance of the tube.

Examination of Eqs. (194) and (195) shows that *in order for the by-passing of the self-bias resistance to be adequate, the voltage developed across the bias impedance Z_c by the amplified plate current must be small compared with the signal voltage being amplified.* There is no difficulty in meeting this requirement at radio and the higher audio frequencies, but at low audio frequencies the capacity required tends to become excessively large. The effect of insufficient by-pass condenser, or of no condenser at all, depends upon the character of the load impedance in the plate circuit, as is apparent from Fig. 261.

When there is an input transformer, the arrangement illustrated in Fig. 262 provides a simple means of avoiding regeneration at the lower audio frequencies. By making the reactance of the condenser C_c much smaller than the resistance R_1, the grid return to alternating current is effectively to the cathode, so that any alternating voltage developed across the bias impedance does not result in a potential being applied

---

use of the equivalent plate circuit of the vacuum tube, which leads to the relation

$$e_r = -\mu e_g \frac{Z_c}{Z_L + R_p + Z_c}$$

The voltage e_g actually acting on the grid hence is

$$e_g = e_s + e_r = e_s - \frac{Z_c}{Z_L + R_p + Z_c}\mu e_g$$

Solving this for e_g/e_s gives

$$\frac{e_g}{e_s} = \frac{1}{1 + \dfrac{\mu Z_c}{Z_L + R_p + Z_c}}$$

Equation (194) now follows at once.

between the grid and cathode. At the same time the circuit arrangement is such that to direct currents the grid is at ground potential while the cathode is positive with respect to ground. By making the resistance R_1

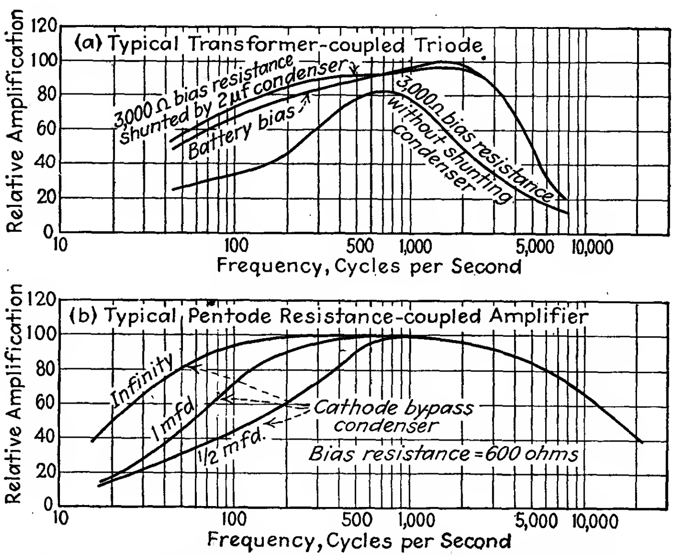


FIG. 261.—Effect of insufficient by-pass capacity in different amplifiers employing self-bias resistors.

large it is possible to meet the requirements for proper operation with a reasonable value of capacity for C_c.

94. Sources of Anode Power.—The commonest sources of direct-current power for operating the anode (*i.e.*, plate, screen-grid, or space-charge-grid) electrodes of vacuum tubes are rectifier-filter systems operating from commercial lighting circuits, 110-volt direct-current lighting circuits, dynamotors and vibrators operating from low-voltage storage batteries, dry batteries, and high-voltage storage batteries. Rectifier-filter systems are most commonly used because of their economy and convenience. Direct-current lighting circuits of 110 volts are employed to a limited extent in equipment designed so that it can be operated in localities where this power is used for lighting purposes. Vibrators and dynamotors are found in automobile radios, aircraft equipment, etc., where the only convenient source of power is a low-voltage storage battery. High-voltage storage batteries are common in commercial radio stations where enough tubes are in operation to provide a heavy direct-

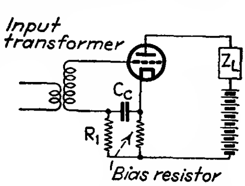


FIG. 262.—Circuit that avoids regeneration from bias resistor when an input transformer is used.

current load. Dry batteries, or "B batteries," are used primarily where other sources of power are either not available or not convenient, as in portable equipment, receivers in isolated locations, etc. Such batteries consist of a number of small dry batteries assembled in a moistureproof carton.

95. Rectifiers for Supplying Anode Power.—As a result of a long period of evolution and development the high-vacuum thermionic rectifier and the hot-cathode mercury-vapor rectifier have become practically standard in vacuum-tube power-supply systems. Mercury-arc rectifiers are, however, used in some installations requiring very large amounts of direct-current power, while copper oxide rectifiers find application for such purposes as supplying direct current for speaker fields.

High-vacuum Thermionic Rectifiers.—The high-vacuum thermionic rectifier consists of a vacuum tube containing an electron-emitting cathode surrounded by an anode or plate electrode. Such a two-element tube acts as a rectifier because it will pass current only when the plate is positive with respect to the cathode, and so when placed in series with an alternating supply voltage and a load impedance will permit current to flow in only one direction.

The characteristics of the two-electrode vacuum tubes were discussed in detail in Sec. 27 and are incorporated in the curves giving the relationship between the anode current and anode voltage, such as those of Fig. 51. When the plate voltage is not too large, the electrons are emitted from the cathode more rapidly than they can be drawn to the positive plate, causing a space charge to be formed in the vicinity of the cathode and making the anode current substantially independent of the cathode temperature. Under these conditions the current that the anode draws from each part of the filament is proportional to the $3/2$ power of the anode voltage with respect to that part of the filament, and when the voltage drop in the filament is negligible in comparison with the anode voltage, which is nearly always the case in high-vacuum rectifiers, the total anode current is almost exactly proportional to the $3/2$ of the anode voltage. If the anode voltage is very high, however, the electrons are drawn away from the cathode as fast as emitted. The plate current is then determined only by the electron emission and is independent of the plate voltage, *i.e.*, voltage saturation is present.

The important characteristics of the high-vacuum thermionic rectifier are the allowable peak plate current and the maximum allowable peak inverse voltage. The peak plate current represents the maximum electron emission which the cathode can be counted upon to supply during the useful life of the tube and is therefore determined by the cathode. Since the rectifier never allows current to flow for more than half of the

time, the average plate current, *i.e.*, the d-c output current, will never exceed one-half of the peak plate current and may be less. The maximum allowable inverse plate voltage is the largest negative voltage that may be applied to the plate with safety, and determines the direct-current voltage that can be obtained from the rectifier tube. The exact relationship between direct-current output voltage and the allowable inverse voltage depends upon the rectifier circuit employed, but in general the inverse voltage will be at least as great as the direct-current voltage and in certain rectifier connections will be π times as great.

High-vacuum thermionic rectifier tubes are constructed in much the same way as the corresponding power tubes; in fact, most types of rectifier tubes are merely standard filament-type three-electrode tubes with the grid omitted. The only exception to this is in the case of small rectifiers used in supplying anode power for radio receivers, where the low inverse voltages encountered permit a construction that places the plate very close to the filament. Large rectifiers are water cooled and employ tungsten filaments just as do the water-cooled power tubes, while the medium and small rectifiers are air cooled and make use of either thoriated-tungsten or oxide-coated cathodes. The size of the filament is determined by the required maximum peak plate current, while the spacing of the electrodes and degree of vacuum determine the maximum safe inverse voltage. The losses which must be dissipated by the tube consist of the filament-heating power and the average plate loss. The average plate power will never exceed one-half the instantaneous power that is dissipated in the tube when the voltage drop between plate and filament has the value required to make the plate current equal the allowable peak value, and is less than the plate loss in the corresponding power tube because, when there is no grid to shield the plate, the full space current is obtained with a relatively low plate potential. The characteristics of a number of representative high-vacuum thermionic rectifiers are shown in Table XI.

The high-vacuum thermionic rectifier can be built to withstand inverse voltages in excess of 100,000, and commercial tubes having peak plate currents of 7.5 amp. are available. The efficiency of the high-vacuum thermionic rectifier is high, particularly when used to develop large direct-current voltages, for the voltage drop in the tube is a relatively small fraction of the output voltage, and the cathode-heating power is only a small fraction of the output. The development of the hot-cathode mercury-vapor tube has, however, limited the field of the high-vacuum thermionic rectifier to the production of direct-current voltages greater than those which can conveniently be obtained from the mercury-vapor type of tube, and to low-power low-voltage applications (notably in broadcast receivers) where the superior ruggedness and

freedom from transients of the high-vacuum type of tube makes it more satisfactory than the hot-cathode mercury-vapor rectifier.

TABLE XI.—CHARACTERISTICS OF TYPICAL HIGH-VACUUM THERMIONIC RECTIFIER TUBES

Rating				Filament data			
Type	Maximum allowable peak plate current, milli-amperes (approx.)	Maximum safe inverse voltage	Voltage drop with one-half peak current	Volts	Amperes	Watts	Type
80*	250	1,400	62	5	2	10	Oxide filament
5Z3*	500	1,400	61	5	3	15	Oxide filament
83V*	400	1,100	22	5	2	10	Heater
84*	120	1,000	17	6.3	0.5	3.2	Heater
81	340	2,000	120	7.5	1.25	9.4	Oxide filament
217A	600	3,500	210	10	3.25	32.5	Thoriated filament
836	1,000†	5,000	110	2.5	5	12.5	Heater
214‡	7,500	50,000	2,000	22.0	52.0	1,144	Tungsten filament

* These tubes have two cathodes and two anodes, and so are essentially two half-wave rectifiers in one envelope. The allowable plate current and voltage drop are given for a single anode, but the filament data are for both filaments.

† The maximum allowable *average* plate current in this tube is 0.25 amp.

‡ This is a water-cooled tube now becoming obsolete.

The Hot-cathode Mercury-vapor Rectifier.[1]—The hot-cathode mercury-vapor rectifier is essentially a high-vacuum thermionic rectifier which contains mercury vapor in equilibrium with liquid mercury. The distinguishing characteristic of such a tube is that the plate is able to draw the full electron emission of the cathode when only 15 to 20 volts positive with respect to the cathode. The reasons for this behavior can be explained as follows: When the plate is more positive than the filament, some of the electrons collide with mercury molecules, and, if the voltage

[1] For a more detailed discussion of tubes of this type see H. C. Steiner and H. T. Maser, Hot-cathode Mercury-vapor Rectifier Tubes, *Proc. I.R.E.*, vol. 18, p. 67, January, 1930; H. C. Steiner, Hot Cathode Mercury Rectifier Tubes for High Power Broadcast Transmitters, *Proc. I.R.E.*, vol. 23, p. 103, February, 1935.

difference between plate and cathode is at least 10.4 volts, which is the ionization potential of mercury, some of these collisions will be sufficiently severe to knock electrons out of the mercury-vapor molecules (*i.e.*, will cause ionization by collision). When the plate becomes about 15 volts positive with respect to the cathode, these ionizing collisions become very numerous. The electrons thus produced are attracted to the plate, but because of the low pressure of the mercury vapor (1 to 30 μ) the increase in plate current which results from the ionization is entirely negligible. The positive mercury ions which result from mercury molecules losing electrons are drawn toward the filament, but, being very heavy, they move much more slowly than do the electrons (only about $\frac{1}{600}$ as fast), so that, although the rate of production of positive ions is relatively slow compared with the rate of emission of electrons from the filament, the number of positive ions that is present at any one time in the interelectrode space is of the same order of magnitude as the number of electrons. The positive mercury ions therefore produce a positive space charge that is of about the same magnitude as the negative space charge produced by the electrons emitted from the filament, giving a resultant space charge that approaches zero. This neutralization of the negative space charge by the slow-moving ions permits the plate to draw the electrons from the cathode as fast as they are emitted when the plate is only 15 to 20 volts positive with respect to the cathode. The result is a relation between anode voltage and anode current such as illustrated in Fig. 85 and also discussed in Sec. 38.

Since the positive ions eventually fall into the filament, it might be thought that the filament life of hot-cathode mercury-vapor tubes would be very short. Experiments by Langmuir and Hull have shown, however, that cathode disintegration by positive-ion bombardment with mercury molecules does not take place to an appreciable extent unless the positive mercury ions have fallen through a potential that is in excess of about 22 volts. *If the plate is therefore never allowed to become more than 22 volts positive with respect to the filament, the positive-ion bombardment of the filament produces no injurious effect.* It is therefore apparent that, at plate voltages of 15 and 20 volts positive with respect to the cathode, it is possible to produce sufficient ionization to enable the plate to draw the entire electron emission of the cathode while at the same time avoiding cathode disintegration by positive-ion bombardment.

The important characteristics of the hot-cathode mercury-vapor rectifier are the maximum allowable peak plate current, the maximum permissible average plate current, and the maximum safe inverse plate voltage. The peak plate current is determined by the electron emission that can be obtained from the filament and is unaffected by the presence of the mercury vapor. The safe average current is determined by the

allowable heating of the plate. The maximum safe inverse plate voltage is the sparking voltage through the low-pressure mercury vapor, and is somewhat lower than would be the case with the mercury removed.

More care must be taken in the operation of hot-cathode mercury-vapor rectifiers than is necessary with high-vacuum tubes. In the first place the bulb temperature must be maintained between definite limits because this temperature determines the pressure of the mercury vapor. If the pressure is too low (low operating temperature), the intensity of ionization is reduced to the point where the voltage drop across the tube that is needed to cause neutralization of the negative space charge exceeds the cathode disintegration value, while, when the pressure is high (high operating temperature), the inverse voltage at which spark-over or flashback takes place is reduced excessively. These effects of temperature are shown in Fig. 263. In the second place the instantaneous plate current must never be permitted even momentarily to exceed the allowable peak plate current, since in increasing the plate current there is danger that the voltage drop in the tube will reach the value at which cathode disintegration starts. Thus a momentary short circuit which would merely heat up the plate of a high-vacuum tube will cause permanent damage to a mercury-vapor tube. Finally, the filament of a hot-cathode mercury-vapor tube must be brought to full operating temperature before the plate voltage is applied, for otherwise the voltage drop in the tube during the warming-up process will exceed the cathode-disintegration value and the filament will be permanently damaged. Where mercury-vapor types of rectifiers are employed, it is customary to insert a time-delay relay in series with the rectifier plate circuit, which is thereby held open until the filament has had time to reach its operating temperature.[1]

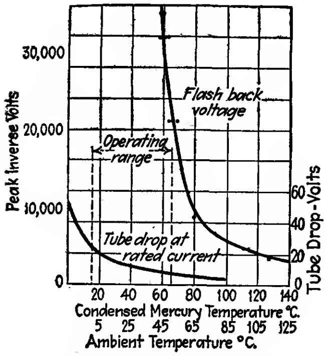


FIG. 263.—Effect of operating temperature on voltage drop and flashback voltage of a hot-cathode mercury-vapor tube. The difference between the mercury and the ambient (room) temperature represents the temperature rise of the coldest part of the tube and is determined by the design.

In connecting hot-cathode mercury-vapor rectifier tubes in parallel in order to obtain high output currents, the anodes of the two tubes should be connected to opposite ends of an inductance, to the center of which is brought the line connection. This arrangement is necessary to

[1] In very small hot-cathode mercury-vapor tubes the filament-heating time is so small that it is common practice to turn on the filament and plate voltage simultaneously.

insure that the two tubes will carry equal anode currents. Otherwise the tube having the smallest internal voltage drop would usually carry nearly all the current even when the difference in the voltage drops required to produce copious ionization in the tubes is only a fraction of a volt.

The distinctive constructional features of a hot-cathode mercury-vapor tube are the glass envelope, relatively small plate, and an oxide-coated cathode having a filament voltage never over 5 volts. The glass envelope and small plate can be used even in tubes having the highest anode current ratings because of the low power loss in the tube. The

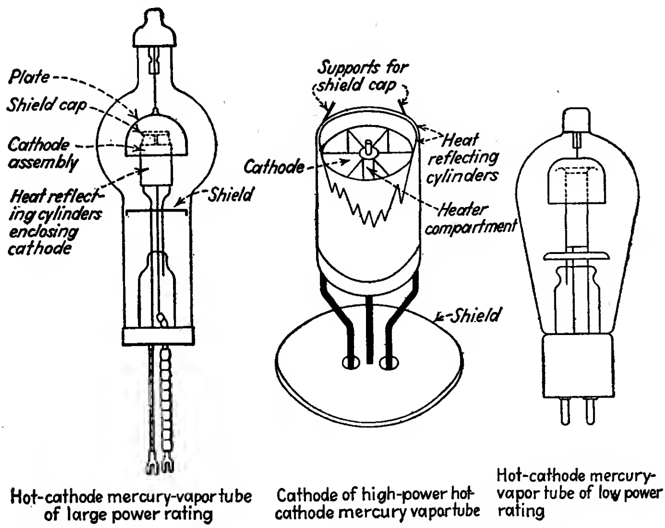


FIG. 264.—Typical hot-cathode mercury-vapor tube, together with details of the heater type cathode.

plate is usually in the form of a cup fitting over the cathode, as in Fig. 264, because this reduces the tendency to flash back, and also shields the plate-cathode space from external electrostatic fields. Oxide-coated cathodes are used because of their high efficiency, while low-filament voltages are necessary because the total voltage drop in the tube is normally about 15 volts and the filament potential should be small compared with this. In hot-cathode mercury-vapor rectifier tubes having high current ratings, the cathodes are usually of the heater type, constructed as shown in Fig. 264*a* in the form of a cup containing vanes or disks that are coated with emitting oxide. This provides a large emitting surface in proportion to heat-radiating area and is permissible because the space-charge neutralizing action of the positive ions penetrates even into the remote pockets of the cathode. Such a cathode is heated from a central heater and is in turn surrounded by one or more shields which are polished to reduce the radiation of heat.

Compared with the high-vacuum rectifier the hot-cathode mercury-vapor tube has the advantage of higher efficiency, better regulation, lower filament power, and low first cost. At the same time the mercury-vapor tube has a limited inverse-voltage rating, displays a tendency to flash back, produces radio-frequency transients, and will suffer damage to the cathode as a result of momentary overloads. When these considerations are balanced together, the result is that the hot-cathode mercury-vapor tube is found to be best for use with transmitters, where the large powers involved make economic considerations important, but not in radio receivers where ruggedness and freedom from radio-frequency transients are fundamental considerations.

TABLE XII.—CHARACTERISTICS OF TYPICAL HOT-CATHODE MERCURY-VAPOR TUBES

Type	Rating			Filament data		
	Maximum allowable peak plate current, amperes	Maximum allowable average plate current, amperes	Maximum safe inverse voltage	Volts	Amperes	Watts
866	1.0	0.250	7,500	2.5	5	12.5
866A	1.0	0.250	10,000	2.5	5	12.5
872	5.0	1.25	7,500	5	10	50
872A	5.0	1.25	10,000	5	6.75	33.75
869A	10.0	2.5	20,000	5	18	90
857	40.0	10	22,000	5	30	150
870	450	75	16,000	5	65	325

Mercury-arc Rectifiers.—Mercury-arc rectifiers are used to provide direct-current power in some of the very large radio transmitters. The rectifiers used for this purpose are of the grid-controlled type developed for industrial use, and give an output voltage that can be varied by changing the phasing of the grid electrodes. Compared with a hot-cathode mercury-vapor rectifier, the mercury arc has the advantage of lower operating cost, since there are no tube replacements, and a peak current not limited by the emission capabilities of a filament. As a consequence of these factors a mercury-arc rectifier competes on more or less even terms with the hot-cathode mercury-vapor rectifier in installations requiring power of the order of 100 kw and more.[1]

[1] A description of a typical grid-controlled mercury-arc rectifier suitable for radio use is described by S. R. Durand, Steel Cylinder Grid Controlled Mercury-arc Rectifiers in Radio Service, *Proc. I.R.E.*, vol. 23, p. 372, April, 1935.

The Copper Oxide Rectifier.—The copper oxide rectifier makes use of the fact that, when a thin film of cuprous oxide is formed upon a metallic copper surface, the resistance that this film offers to electrical currents is small for currents flowing in one direction and high for currents going the opposite way. Each individual copper oxide film will stand a back voltage of only a few volts, so that it is ordinarily necessary to employ a number of films in series. This is accomplished by building up a pile of alternate rectifying and lead washers, as shown in Fig. 265, and clamping the whole together. Rectifiers of the copper oxide-film type find their greatest usefulness for such low-voltage high-current uses as charging batteries and developing direct-current power for speaker fields, operating control equipment, etc.[1]

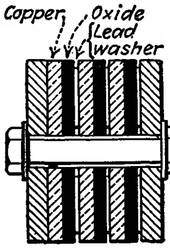


FIG. 265.—Assembly of units of copper oxide type of thin-film rectifier. The cuprous oxide film is formed directly on the surface of the copper and has its thickness exaggerated in the figure.

96. Rectifier Circuits.—The various types of rectifier connections that may be employed with a single-phase source of power are shown in Fig. 266, together with the wave form of the voltage which is developed across a resistance load.

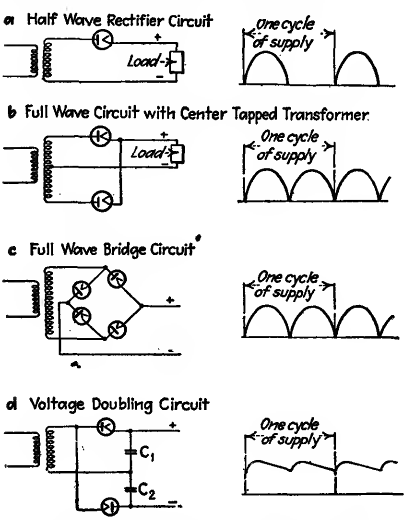


FIG. 266.—Rectifier circuits for operation with single-phase power sources, together with wave forms of voltage developed across a resistance load.

The circuit shown at Fig. 266*a*, in which a single rectifier is placed in series with the source of alternating voltage and the load impedance, is called a half-wave rectifier circuit. It has the very great disadvantage of delivering an output voltage that is far from being a continuous direct-current potential and of producing a direct-current magnetization in the core of the supply transformer as a result of the rectified current that flows through the secondary; it is therefore seldom used.

The rectifier circuit most commonly employed with a single-phase power source is shown in Fig. 266*b*, and consists of two rectifier units operating in conjunction with a center-tapped transformer in such a way

[1] An excellent discussion of the theory of copper-oxide rectifiers is given by J. Slepian, Thin Film Rectifiers, *Trans. Amer. Electrochem. Soc.*, vol. 54, p. 201, 1928. Information on the performance of such rectifiers is given by L. O. Grondahl and P. H. Geiger, A New Electronic Rectifier, *Trans. A.I.E.E.*, vol. 46, p. 357, 1927.

that the two tubes alternately supply rectified current to the load, giving the so-called "full-wave" rectifier output voltage shown. Such a full-wave rectifier not only produces a direct-current voltage that is more nearly constant than that produced by the half-wave rectifier, but also avoids direct-current saturation in the core of the supply transformer since the direct-current magnetizations in the two halves of the transformer secondary are opposed to each other and give zero resultant magnetization.

The bridge-type of full-wave rectifier shown at Fig. 266*c* requires four rectifier units instead of the two called for by the center-tapped transformer arrangement, but has the advantage of requiring only one secondary winding instead of two. This circuit is widely used with copper-oxide rectifiers but is seldom employed with rectifiers of the thermionic type since the different cathodes are not at the same potential and so cannot be connected in parallel and supplied from a single filament transformer secondary.

The rectifier circuit shown at Fig. 266*d* is known as the voltage-doubling circuit because it has the unique property of delivering a direct-current voltage that approaches twice the alternating voltage which is supplied by the transformer. This is accomplished because of the fact that the two condensers C_1 and C_2 are alternately charged to the full voltage of the transformer, and, since the condensers are in series, the output voltage can reach twice that of the alternating supply. The wave form of the output delivered by such a rectifier system depends upon the load and the size of condensers C_1 and C_2. The voltage-doubling circuit finds its chief usefulness where the required direct-current output voltage is greater than can be obtained from a single tube, as in x-ray work.

Polyphase Circuits.—When a polyphase source of alternating power is employed, the number of possible rectifier connections is almost unlimited, although only a relatively few of these are of practical importance.[1] The polyphase rectifier circuits most commonly used with three-phase power sources are shown in Fig. 267 and develop across a resistance load voltages that have the wave forms indicated in the figure. The three-phase half-wave circuit is essentially three half-wave rectifiers of the type shown in Fig. 266*a*, with each leg of the secondary Y forming one phase. In such an arrangement each rectifier tube carries current one-third of the time, and the output wave pulsates at three times the frequency of the alternating-current supply. In order to avoid direct-

[1] For a more complete list of polyphase rectifier circuits see R. W. Armstrong, Polyphase Rectification Special Connections, *Proc. I.R.E.*, vol. 19, p. 78, January, 1931. Also see D. C. Prince and F. B. Vodges, "Mercury-arc Rectifiers and Their Circuits," McGraw-Hill Book Company, Inc., New York.

current saturation in the transformer, it is necessary to employ a three-phase transformer rather than three single-phase transformers.

The circuit of Fig. 267*b*, which employs six tubes and two three-phase Y-connected secondaries, is essentially two three-phase half-wave rectifiers of the type shown in Fig. 267*a* connected in parallel, but with the polarities such that during the period when the output voltage of one three-phase unit is at a minimum the output of the other unit is maximum,

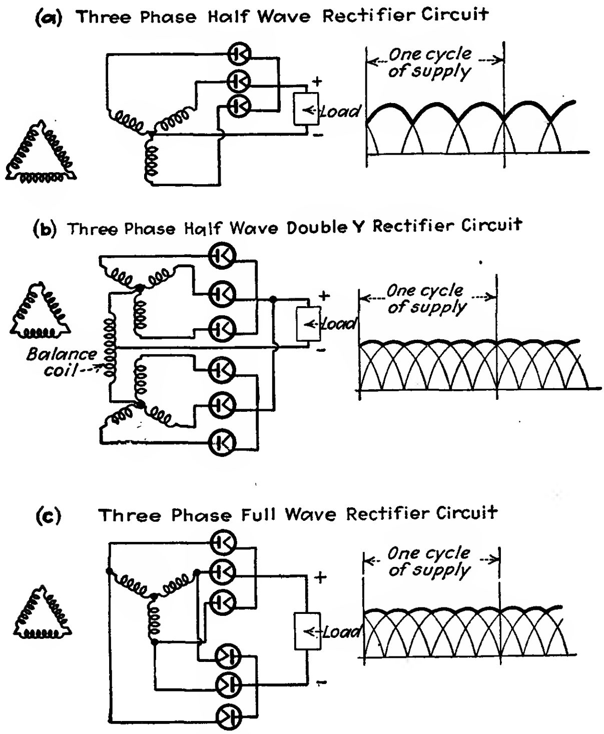


FIG. 267.—Rectifier circuits for operation with three-phase power sources, together with wave forms of voltage developed across resistance load.

so that the ripple in the output wave is small and has a fundamental frequency six times that of the power supply. The two three-phase units are connected in parallel through an interphase reactor (or "balance coil") which enables each three-phase unit to operate independently. If it were not for this reactor, each tube would carry the load current only one-sixth of the time, whereas with the reactor each tube carries current one-third of the time, and at any instant there are always two tubes delivering current to the load. The balance coil should have sufficient inductance so that the alternating current flowing as a result of the voltage that exists across the coil has a peak value less than one-half

the normal direct-current load current (*i.e.*, a peak value less than the direct current in one leg). Since the direct current flows in opposite directions in the two halves of the winding, no direct-current saturation is present, and an air gap need not be provided in the core. The result is that the interphase reactor requires only a small amount of material.

The three-phase full-wave rectifier circuit shown at Fig. 267*c* gives the same output wave as does the double three-phase half-wave rectifier of Fig. 267*b* but differs in that the tubes are arranged so that full-wave rectification is obtained through each leg of the secondary winding. This circuit requires only one three-phase secondary and no interphase reactor, but the filament transformer must have four separate secondaries.

Polyphase rectifiers are used where the direct-current power required is in the order of 1 kw or more. Compared with the single-phase circuits, the polyphase rectifiers, particularly those of the full-wave type shown in Fig. 267*b* and *c*, develop an output voltage wave that is much closer to a steady direct-current potential than is the case with single-phase arrangements, and the more desirable polyphase circuits give a higher output voltage in proportion to the peak inverse voltage and also utilize the possibilities of the transformer more effectively.

Filament Transformers.—The filaments of thermionic rectifiers are usually heated by alternating current obtained from a filament transformer having a secondary, or secondaries, insulated to withstand the direct-current output voltage. In the case of small rectifiers, such as those employed in radio receivers and low-power transmitters, it is common practice to add a special secondary to the rectifier transformer for the purpose of securing filament power, but where appreciable amounts of power are involved it is preferable to employ a separate filament transformer. The number of secondaries required for filament heating ranges from one in the case of the simpler circuits to four for the three-phase full-wave circuit, where it will be noted one of the four secondaries must have sufficient capacity to operate three filaments, while the other three each supply only one filament. The filament windings are usually provided with a center tap to which the cathode connection of the rectifier is brought in order to equalize the flow of plate current in the two legs of the filament.

In certain cases, notably "a-c-d-c receivers," and in power-supply systems employing a storage battery and vibrator, heater-type rectifier tubes are employed and are designed to have sufficient insulation between heater and cathode to permit operation of the heater from a grounded source.

97. Filter Circuits Having a Series-inductance Input.—The pulsating voltage delivered by the rectifier output can be smoothed into a steady direct-current voltage suitable for applying to the anode circuit of a

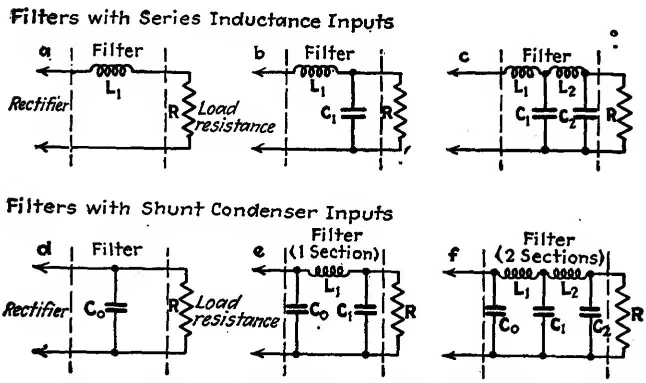


FIG. 268.—The filter circuits most commonly employed to smooth out the pulsating rectifier output into a steady direct-current voltage.

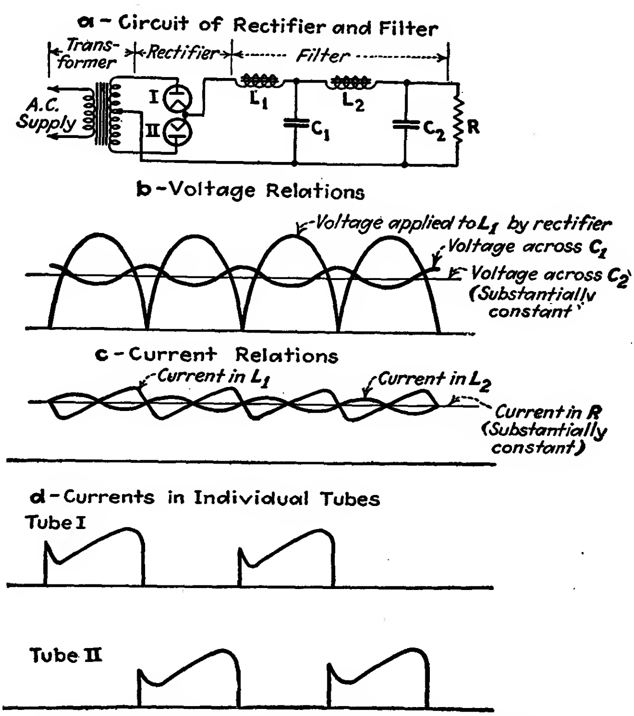


FIG. 269.—Oscillograms showing action taking place in filter having a series-inductance input when supplied power from a single-phase full-wave rectifier. These curves are idealized in that they neglect transformer leakage reactance, tube drop, and the effect of energy losses in the filter inductances and condensers.

vacuum tube by being passed through an electrical network, commonly called a filter, which ordinarily consists of series inductances and shunt condensers. The most commonly used filter circuits are those shown in Fig. 268 and may be divided into two general classes according to whether the input consists of a series inductance or a shunt condenser. Each type of filter may be further subdivided according to the number of sections or filter elements involved.

The action that takes place in a properly adjusted filter having a series-inductance input can be understood from an examination of the oscillograms of Figs. 269 and 270. Consider first the case of a full-wave single-phase rectifier delivering its output to the filter of Fig. 268c. The current flowing into the filter tends to increase when the voltage output of the rectifier is high and tends to decrease when the rectifier voltage is low, but, if the input inductance is reasonably large, as is the case in Fig. 269, these variations in current are relatively small, and to a first approximation the input current can be considered constant, with each rectifier tube carrying the full current for one-half of the time.[1] The first condenser tends to absorb what variations there are in the current entering the input inductance, with the result that the voltage that appears across the terminals of this condenser is more nearly constant than is the input current.

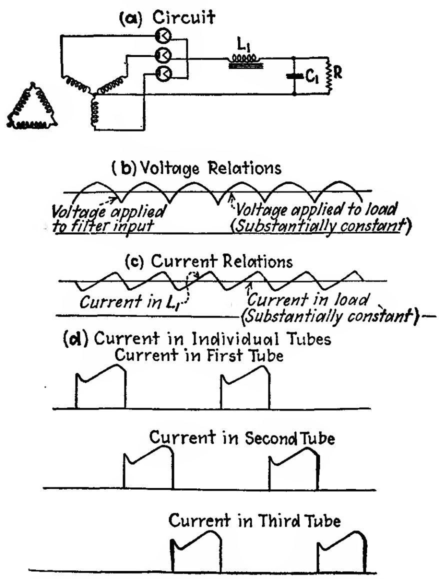


Fig. 270.—Oscillograms showing action taking place in a filter having a series-inductance input when supplied power from a three-phase half-wave rectifier. These curves neglect the effect of transformer reactance, tube voltage drop, and losses in the filter elements.

The voltage across the first condenser is applied to the second inductance and is smoothed out still more by the action of this inductance and the second condenser, with the result that the potential appearing across the load is substantially pure direct current with a value equal to the average output voltage of the rectifier.

The action that takes place in the three-phase half-wave rectifier circuit is somewhat similar and is shown in Fig. 270. The special

[1] In the case of the full-wave bridge circuit there are always two rectifier tubes in series so that each tube carries the full current half of the time even though there are two tubes operating at any instant.

features to note here are that the output wave of this type of rectifier is more nearly constant than that of the full-wave single-phase type, and that each tube carries the output current only one-third of the time.

Examination of the oscillograms of Figs. 269 and 270 shows that the voltage across the first condenser C_1 is nearly constant at a value corresponding to the average voltage of the rectifier output. The potential difference across the first filter inductance hence approximates the difference between the actual instantaneous voltage of the rectifier output and the average value of this output voltage. When the voltage output of the rectifier is above the average, the current tends to increase, while when it is less than the average the current through the inductance decreases. The amplitude of the current variation that results depends upon the size of the inductance in the way that is shown in Fig. 271. When the input inductance is very large, the current variations are negligibly small, as shown by curve a, and each tube passes a wave of current that is approximately square. If the input inductance is only moderately large, the input current varies as shown at b, and the current waves passed by the individual tubes have a ratio of average to peak value that is lower than in case a. As the input inductance is reduced, the current variations become larger until a point is finally reached when the rectifier ceases to draw current continuously throughout the cycle, which results in the condition shown at Fig. 271c. This last condition is to be avoided under normal operating conditions because it leads to a low ratio of average to peak anode current in the individual rectifier tubes and also results in poor regulation of the direct-current potential.

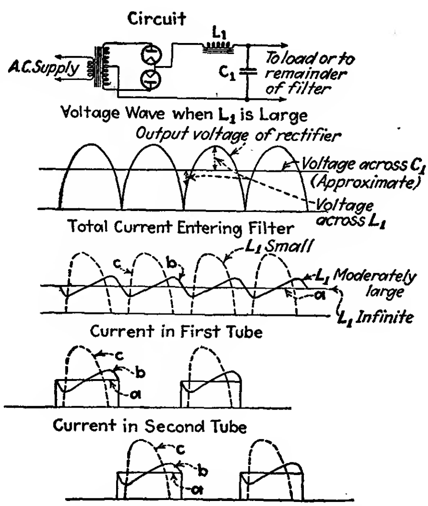


FIG. 271.—Oscillograms showing the effect of changing the size of the series input inductance when a single-phase full-wave rectifier is used.

Analysis of Voltage Delivered by Rectifier to the Filter.—The action taking place in a filter having an input inductance of sufficient size to maintain a continuous flow of current from the rectifier can be calculated with an accuracy sufficient for most practical purposes by considering that the rectifier applies to the filter input a voltage having a wave shape shown by the idealized curves of Figs. 266 and 267. This neglects the resistance and the leakage reactance of the transformer and the drop

in the rectifier but is justified because these factors are merely modifying influences in rectifier-filter systems of the type used in supplying anode power. The idealized output wave of the rectifier can be considered as consisting of a direct-current component upon which are superimposed alternating-current voltages. Thus, in the case of the full-wave single-phase rectifier, the output wave has the equation

Output voltage of single-phase full-wave rectifier =

$$\frac{2E}{\pi}\left(1 - \frac{2}{3}\cos 2\omega t - \frac{2}{15}\cos 4\omega t - \frac{2}{35}\cos 6\omega t - \cdots\right) \tag{196}$$

where E represents the crest value of the alternating-current voltage applied to the rectifier tube and ω is the angular velocity ($2\pi f$) of the supply frequency. In this case the direct-current component of the

TABLE XIII.—CHARACTERISTICS OF RECTIFIERS OPERATED WITH A FILTER SYSTEM HAVING A SERIES INPUT INDUCTANCE

	Rectifier circuit				
	Single-phase, full-wave, center-tapped connection	Single-phase, full-wave bridge	Three-phase, half-wave	Double three-phase, half-wave	Three-phase, full-wave
Voltage Relations (Direct-current component of output voltage taken as 1.0):					
a. R.m.s. value of transformer secondary voltage (per leg)	1.11*	1.11	0.855	0.855	0.428
b. Maximum inverse voltage	3.14	1.57	2.09	2.42	1.05
c. Lowest frequency in rectifier output (F = frequency of power supply)	$2F$	$2F$	$3F$	$6F$	$6F$
d. Peak value of first three alternating-current components of rectifier output					
Ripple frequency	0.667	0.667	0.250	0.057†	0.057
Second harmonic of ripple frequency	0.133	0.133	0.057	0.014	0.014
Third harmonic of ripple frequency	0.057	0.057	0.025	0.006	0.006
Current Relations:					
e. $\dfrac{\text{Average anode current}}{\text{Peak anode current}}$	0.500	0.500	0.333	0.333	0.333
f. $\dfrac{\text{Average current per anode}}{\text{Direct-current load current}}$	0.500	0.500	0.333	0.167	0.333
Transformer Utilization Factors:					
g. Primary	0.900	0.900	0.827	0.955	0.955
h. Secondary	0.637	0.900	0.675	0.675	0.955

NOTE: This table assumes that the input inductance is sufficiently large to maintain the output current substantially constant, and neglects the effects of voltage drop in the rectifier and the transformers.

* Secondary voltage on one side of center tap.

† The principal component of the voltage across the balance coil has a frequency of $3F$ and a peak amplitude of 0.500.

output wave is $2/\pi$ times the crest value of the alternating-current wave, the lowest frequency alternating-current component in the output is twice the supply frequency and has a magnitude that is two-thirds of the direct-current component, and the remaining alternating-current components are harmonics of this lowest frequency component. Table XIII gives results of such analyses for the waves delivered by the single-phase full-wave rectifier, by the three-phase half-wave rectifier, and by the three-phase full-wave rectifier. It will be observed that in the three-phase half-wave rectifier the lowest alternating-current frequency is three times the frequency of the power supply, while in the three-phase full-wave rectifier it is six times that of the power supply. In all cases the amplitude of the alternating-current components diminishes rapidly as the order of the harmonic is increased.

Calculation of Direct-current and Alternating-current Components of Filter Output.—If the voltage drop in the rectifier and the leakage reactance of the supply transformer are neglected, the direct-current voltage delivered to the load is less than the direct-current input to the filter, as calculated from Table XIII, by an amount equal to the voltage drop in the resistance of the filter inductances, and the regulation of the output voltage is then the regulation that would result from the filter resistances. Actually, the resistance and the leakage reactance of the supply transformer and the voltage drop in the rectifier tube cause the direct-current output voltage to be lowered somewhat and make the regulation poorer, although Table XIII will give the magnitude of the direct-current output voltage delivered to the filter input with an accuracy sufficient for most purposes.

The alternating-current voltage that appears across the output of the filter is the potential that is developed across the filter output when the alternating-current voltages given in Table XIII are applied to the filter input. Since the smoothing action of the filter results from the fact that the series inductances of the filter choke out these alternating-current voltages, while the shunt condensers tend to short-circuit them, the output condenser must have a reactance that is low compared with the load resistance, while each inductance must have a high reactance compared with the reactance of the condenser that immediately follows it. Furthermore the input inductance must also have sufficient reactance in relation to load resistance to satisfy Eq. (198) if current is to flow into the filter throughout the cycle.

An exact determination of the alternating-current voltage that appears across the output of the filter involves considerable labor because of the complicated electrical networks involved, but for most purposes it is permissible to simplify the analysis by assuming that the reactance of each condenser is small compared with the reactance of the inductances

immediately preceding and following the condenser, and that the reactance of the output condenser is small compared with the load resistance. The fraction of the alternating-current voltage applied to the filter input that reaches the filter output is then given by the following equations.[1]

For the filter of Fig. 268*c*:

$$\frac{\text{Alternating-current voltage across load}}{\text{Alternating-current voltage applied to input}} = \frac{1}{\omega^4 L_1 L_2 C_1 C_2} \tag{197a}$$

For the filter of Fig. 268*b*:

$$\frac{\text{Alternating-current voltage across load}}{\text{Alternating-current voltage applied to input}} = \frac{1}{\omega^2 L_1 C_1} \tag{197b}$$

For the filter of Fig. 268*a*:

$$\frac{\text{Alternating-current voltage across load}}{\text{Alternating-current voltage applied to input}} = \frac{R}{\sqrt{R^2 + (\omega L_1)^2}} \tag{197c}$$

In these equations ω is the angular velocity ($2\pi f$) corresponding to the frequency of the component involved, and the alternating-current voltage applied to the input is given by Table XIII for different rectifier connections. An examination of Eq. (197) shows that the filtering action increases very rapidly with the number of filter elements, *i.e.*, the number of inductances and capacities. The filter is also seen to be more effective the higher the frequency, and this, coupled with the fact that the largest component of the alternating voltage in a rectifier output is always the one having the lowest frequency, makes it permissible to neglect all frequency components in the rectifier output except the fundamental in calculating the alternating-current voltage that will appear across the load.

Factors Involved in the Design of Rectifier-filter Systems.—The input inductance of a filter should ordinarily be of sufficient size to maintain a continuous flow of current from the rectifier under normal operating conditions, *i.e.*, the crest or peak value of the alternating current flowing in the input inductance should be less than the direct-current load current. This condition is realized by satisfying the approximate relation[2]

[1] The derivation of a relation such as given by Eq. (197*a*) is as follows: The alternating current which flows in L_1 as a result of an applied voltage E is $E/\omega L_1$ if the reactance of C_1 is small compared with that of L_1. Practically all this current flows through C_1 because of the much higher reactance of the path through L_2, so that the voltage developed across C_1 is $E/\omega^2 L_1 C_1$. This potential is then applied to L_2 and C_2, and, if the reactance of C_2 is low compared with the load resistance, a repetition of the method used to get the voltage across C_1 shows that the alternating-current potential across C_2 is $E/\omega^4 L_1 L_2 C_1 C_2$, and Eq. (197*a*) follows at once.

[2] In order to insure a continuous flow of current in the rectifier, the crest value of the alternating current flowing through the input inductance must not exceed the average or direct current. Since the direct current is equal to the direct-current

$$\frac{\omega L_1}{R_{\text{eff}}} \geq \frac{\text{lowest frequency component of alternating-current voltage in rectifier output}}{\text{direct-current voltage in rectifier output}} \tag{198a}$$

where ωL_1 is the reactance of the input inductance to the lowest ripple frequency and R_{eff} is the effective load resistance, *i.e.*, the actual load resistance plus direct-current resistances of the filter inductances. In the case of the single-phase full-wave rectifier with 60-cycle supply, this becomes

$$L_1 = \frac{R_{\text{eff}}}{1130} \tag{198b}$$

The higher the load resistance, *i.e.*, the lower the direct-current load current, the more difficult it is to maintain a continuous flow of current, and with a given L_1 Eq. (198) will not be satisfied when the load resistance exceeds a critical value.

When the load on an inductance-input filter varies, the voltage regulation will be poor unless Eq. (198) is satisfied at all times. In order to do this with a practical inductance at small load currents, it is necessary to connect a bleeder resistance across the filter output to increase the current in the input inductance. Even then, with the bleeder drawing a reasonable proportion of the total current, the input inductance required will normally be larger than required for suppressing hum. An economical way of meeting this situation is to make the air gap of the input inductance small so that the incremental inductance is high with small currents. With large currents the incremental inductance drops, but Eq. (198) is still satisfied, and the filtering, though not so perfect as with a larger input inductance, is still acceptable. An input inductance adjusted to operate in this way is sometimes referred to as a *swinging choke*.

The ratio of average to peak anode current depends upon the rectifier connection and upon the size of the input inductance. Table XIII gives the results for the common circuits when the input inductance is infinite. When the input inductance is finite, the peak anode current is the sum of the direct-current output current and the crest value of the alternating current flowing in the input inductance. Since the largest part of this alternating current is the component of lowest ripple frequency, the following relation is approximately true, provided Eq. (198) is satisfied:

$$\frac{\text{Peak current with finite input inductance}}{\text{Peak current with infinite input inductance}} = 1 + \frac{E_1}{E_0}\frac{R_{\text{eff}}}{\omega L_1} \tag{199}$$

voltage in the rectifier output divided by R_{eff}, while the crest alternating current in the first inductance is very nearly equal to the fundamental ripple-frequency voltage contained in the rectifier output, divided by the reactance wL_1 of the input inductance to this lowest frequency component of the ripple voltage, Eq. (198a) follows at once.

E_1 is the amplitude of the lowest frequency component of the ripple voltage as given in Table XIII, E_0 is the direct-current output voltage of the rectifier, and R_{eff} and ωL_1 have the same meaning as in Eq. (198*a*).

The maximum inverse voltage that the rectifier tube will be called upon to withstand depends upon the rectifier connections and may vary from only slightly more than the direct-current output up to π times this potential. Values for the commonly used circuits are given in Table XIII.

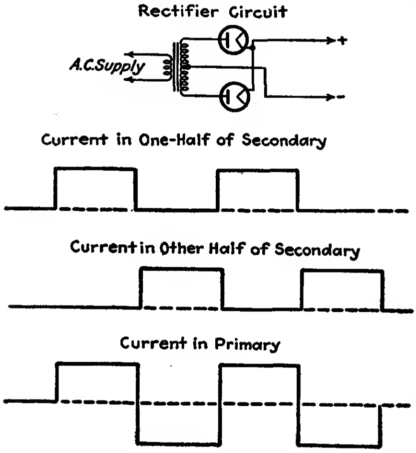


FIG. 272.—Wave shapes of current in primary and secondary windings of a center-tapped transformer supplying a single-phase full-wave rectifier operating into a filter with large series input inductance.

A transformer used to supply power to a rectifier will normally run hotter than when delivering the same amount of energy to a resistance load because of the irregularly shaped current waves drawn by the rectifier. Thus in the case of the full-wave single-phase rectifier the current waves with a large input inductance have the shape shown in Fig. 272. The ratio of direct-current power output to the normal alternating-current rating for the same transformer losses is called the *utilization factor* of the transformer and depends upon the rectifier connections. Table XIII gives the utilization factor of the primary and secondary windings for some of the more commonly used rectifier connections.

By making use of Eqs. (197), (199), and Table XIII it is possible to design rectifier-filter combinations having an input inductance and to calculate everything about the performance except the voltage regulation[1] with an accuracy that is sufficient for most practical purposes. The method by which this is done is illustrated by the following example:

Example.—It is desired to design a three-phase half-wave rectifier-filter system to operate from a 60-cycle power supply and to deliver a direct-current output of 2500 volts and 0.4 amp. with a ripple that must not exceed 2 per cent. If the direct-current resistance of the filter inductances is neglected, the rectifier must deliver a direct-current output voltage of 2500 volts, and Table XIII shows that the r.m.s. voltage which each secondary leg must develop is $2500 \times 0.855 = 2135$ volts. Since the utilization factors of the primary and secondary, as given by Table XIII, are 0.827 and 0.675, respectively, each leg of the primary must have a rating of $2500 \times 0.4/(3 \times 0.827) = 403$ watts, and each leg of the secondary a rating of $2500 \times 0.4/$

[1] Methods have also been devised for calculating the voltage regulation, but these are so complicated as to be beyond the scope of this book. For further information see D. C. Prince and F. B. Vodges, *loc. cit.*

$(3 \times 0.675) = 493$ watts. Table XIII shows that, in order to satisfy Eq. (198), $\omega L_1/R_{eff} > \frac{1}{4}$; since $R_{eff} = 2500/0.4 = 6250$ while $\omega = 2\pi 180$, L_1 must be not less than 1.38 henries. Tentative calculation based on Eq. (197*b*) shows that the filter of Fig. 268*b* with $C_1 = 1.0$ μf and $L_1 = 9.8$ henries will keep the ripple voltage down to 2 per cent and will be generally satisfactory. Reference to Table XIII and Eq. (199) shows that the peak anode current will be $0.4(1 + 0.14) = 0.456$ amp., while the maximum inverse voltage that each rectifier must stand is $2500 \times 2.09 = 5225$ volts. Type 866 mercury-vapor tubes (see Table XII) would meet these requirements. In actual practice the secondary voltage of the transformer would be made

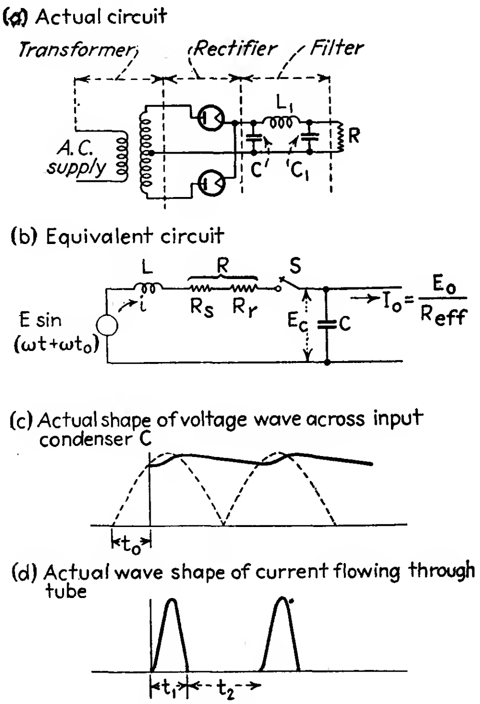


FIG. 273.—Actual and equivalent circuits of condenser-input filter system, together with oscillograms showing operation under typical conditions.

greater than 2135 volts by perhaps 10 per cent to compensate for the loss of voltage caused by the resistance of the filter inductance, the voltage drop in the rectifier, the leakage reactance of the transformer, and the transformer resistance.

98. Filter Circuits Having a Shunt-condenser Input.—When the input to the filter is a shunt condenser, the action is somewhat different from that which takes place when a series inductance is used, as is apparent from the oscillograms of Fig. 273. Each time the crest alternating-current voltage of the transformer is applied to one of the rectifier anodes the input condenser charges up to just slightly less than this peak voltage, and then discharges into the first inductance until another rectifier anode

reaches a peak potential, when the condenser is charged again. The rectifier current flows only a small part of the time, since during most of the cycle the condenser is more positive than all the anodes. During this discharge period the condenser voltage drops off nearly linearly because of the fact that the first filter inductance draws a substantially constant current from the input condenser. The result is that the input condenser applies a saw-toothed voltage wave to the first inductance. This first inductance and the remainder of the filter then act to prevent the alternating-current voltage across the input condenser from reaching the load.

The detailed action that takes place in a condenser-input filter system depends upon the load resistance, the input capacity, the leakage reactance and resistance of the transformer, and the characteristics of the rectifier tube, in a way that is relatively complicated. For purposes of analysis the circuit of Fig. 273*a* can be replaced by the approximate equivalent circuit of Fig. 273*b*. Here the rectifying action of the tube is replaced by a suitable switch S which is assumed to be closed whenever one or the other of the rectifier anodes is conducting current and which is open the remainder of the time. The voltage drop produced in the rectifier tube is accounted for by placing a fixed resistance R_r in the circuit as shown. The transformer is replaced by an equivalent generator having a voltage equal to the open-circuit secondary voltage $E_s \sin \omega t$ measured to the center tap and having an internal impedance equal to the effective leakage inductance L and effective resistance R_s, which are measured between one of the secondary terminals and the center tap when the primary winding is short-circuited. (Note that R_s depends upon both primary and secondary wire resistances.) The input condenser is represented by C, and the first inductance of the filter system is assumed to draw a constant current I_0, which is equal to the direct-current voltage E_0 developed across the input condenser divided by the effective load resistance (*i.e.*, actual load resistance plus resistance of filter inductances).

An analysis of the voltage and current relations existing in the equivalent circuit of Fig. 273*b* gives the results shown in Fig. 274, in which the direct-current voltage across the input condenser, the peak space current through the rectifier tube, and the ripple voltage are expressed as a function of load resistance for various circuit parameters, assuming a 60-cycle source of power and the full-wave rectifier circuit of Fig. 273*a*.[1]

[1] This analysis is carried out as follows: Assume that $E \sin (\omega t + \omega t_0)$ represents the applied voltage, with $t = 0$ being the instant at which the current starts to flow through the rectifier tube (*i.e.*, when the applied voltage in rising just becomes equal to the voltage across the condenser). Then t_0 is the time that elapses between the instant when the applied voltage passes through zero and the instant when the current

Discussion of Properties.—Consideration of Figs. 273 and 274 shows that the addition of a shunt condenser to the input of a filter produces certain fundamental changes in behavior. The output voltage is appreciably higher than with an inductance input while the ripple voltage is less, and the direct-current voltage across the filter input drops as the load current increases instead of being substantially constant. Also the ratio of peak to average anode current in the tube is higher, and the

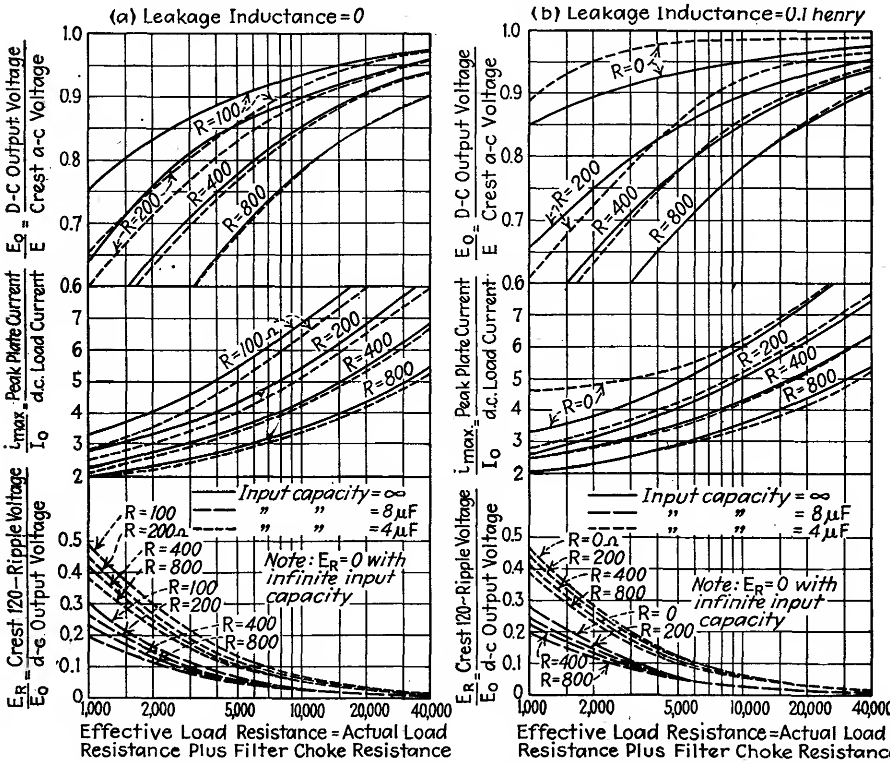


FIG. 274.—Charts giving performance of condenser input filter systems for 60-cycle power source and full-wave rectifier.

utilization factor of the transformer is always rather poor with condenser input systems.

The effect of the circuit parameters upon the performance of the condenser-input system is shown in Fig. 274, while the detailed mechanism

starts to flow. At time $t = 0$ the switch S in Fig. 273*b* is assumed to be closed and to remain closed until a time t_1 when the resulting pulse of current becomes zero, when the switch is assumed to be opened and to remain open until the next half cycle when the second rectifier anode draws current.

During the interval from $t = 0$ to $t = t_1$, during which S can be considered as closed, the sum of the voltage drops in the circuit is equal to the applied voltage,

that produces these results is illustrated in Fig. 275. With the usual circuit proportions the direct-current voltage output and the regulation of this voltage depend primarily upon the load resistance and the source impedance (*i.e.*, impedance formed by the tube and transformer resistance and the transformer leakage inductance) and only to a moderate extent upon the input-condenser capacity. It will be noted from Fig. 274*b* that under certain conditions the leakage inductance resonates with the input capacity to give a higher output voltage with a small input capacity than with infinite input capacity. The ripple voltage is determined primarily by the input capacity and load resistance, and is affected only slightly by the source impedance. The shape of the current pulses flowing through the rectifier tube, and particularly the ratio of peak to average

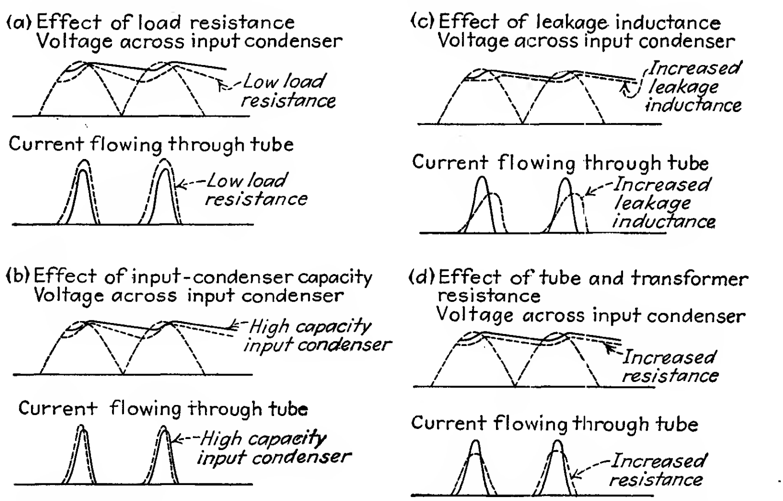


FIG. 275.—Effect of circuit parameters on behavior of condenser-input filter system.

value of these pulses, is controlled by the load resistance and source impedance, but tends to be independent of the input capacity.

It is apparent that, when a high output voltage with good regulation is desired, the source impedance should be made as low as possible. On

---

giving the differential equation

$$E \sin (\omega t + \omega t_0) = L\frac{di}{dt} + Ri + \frac{1}{C}\left[\int_0^t i\,dt - I_0 t\right] + E \sin \omega t_0 \tag{200a}$$

This can be rewritten as:

$$E \sin (\omega t + \omega t_0) - E \sin \omega t_0 + \frac{I_0 t}{C} = L\frac{di}{dt} + Ri + \frac{1}{C}\int_0^t i\,dt \tag{200b}$$

The notation is indicated in Fig. 273*b*. It will be noted that the term $E \sin \omega t_0$ is the voltage across the condenser at the time $t = 0$. Inspection of Eq. (200*b*) shows that this is an ordinary differential equation of an *R-L-C* circuit, with the equivalent of

the other hand, when it is important that the peak currents be kept down, a high source impedance is desirable. A low value of ripple voltage across the input condenser requires the use of a large input capacity. In

---

three separate applied voltages given by the terms on the left-hand side of the equation. Starting with the initial conditions that at $t = 0$, $i = 0$, and $di/dt = 0$, and following the usual method of solving differential equations, the solution is found to be

For oscillatory case $(1/LC > R^2/4L^2)$:

$$i = I_0\left\{1 - \sqrt{1 + \left(\frac{\alpha}{\beta}\right)^2}\epsilon^{-\alpha t} \sin\left[\beta t + \arctan\left(\frac{\beta}{\alpha}\right)\right]\right\} + \frac{E}{Z}\left[\sin(\omega t + \omega t_0 - \gamma) - \sqrt{M^2 + N^2}\epsilon^{-\alpha t} \sin\left(\beta t + \arctan\frac{M}{N}\right)\right] \quad (201a)$$

For non-oscillatory case $(1/LC < R^2/4L^2)$:

$$i = I_0\left[1 - \left(\frac{1}{\alpha_2 - \alpha_1}\right)(\alpha_2\epsilon^{-\alpha_1 t} - \alpha_1\epsilon^{-\alpha_2 t})\right] + \frac{E}{Z}\left\{\sin(\omega t + \omega t_0 - \gamma) - \frac{\sqrt{\omega^2 + \alpha_2^2}}{\alpha_2 - \alpha_1}\epsilon^{-\alpha_1 t} \sin\left(\omega t_0 - \gamma + \arctan\frac{\omega}{\alpha_2}\right) + \frac{\sqrt{\omega^2 + \alpha_1^2}}{\alpha_2 - \alpha_1}\epsilon^{-\alpha_2 t} \sin\left(\omega t_0 - \gamma + \arctan\frac{\omega}{\alpha_1}\right)\right\} \quad (201b)$$

where

i = anode current at time t

I_0 = d-c output current of rectifier

E = crest alternating-current voltage developed by one-half of transformer secondary

R = resistance in series with transformer secondary (see Fig. 273)

L = leakage inductance of one-half of transformer secondary (see Fig. 273)

$X = \omega L - \dfrac{1}{\omega C}$ = reactance of leakage inductance plus input capacity at power-supply frequency

ω = 2π times power-supply frequency

$Z = \sqrt{R^2 + X^2}$

$\gamma = \arctan X/R$

$\beta = \sqrt{\dfrac{1}{LC} - \left(\dfrac{R}{2L}\right)^2}$

$\alpha = R/2L$

$\alpha_1 = \alpha - \sqrt{\alpha^2 - \dfrac{1}{LC}}$

$\alpha_2 = \alpha + \sqrt{\alpha^2 - \dfrac{1}{LC}}$

C = capacity of input condenser

t = time

t_0 = time elapsed between zero of voltage wave and instant when anode current starts to flow (see Fig. 273)

$M = \sin(\omega t_0 - \gamma)$

$N = \dfrac{\sqrt{\omega^2 + \alpha^2}}{\beta} \sin\left(\omega t_0 - \gamma + \arctan\dfrac{\omega}{\alpha}\right)$

Equations (201*a*) and (201*b*) give the current that flows through the rectifier tube from the instant t_0 at which the current begins to flow until some instant t_1 at

considering source impedance, it makes little difference to a first approximation whether the impedance results from high resistance in tube and transformer or from high leakage inductance, as the effects of both are similar.

When hot-cathode mercury-vapor tubes are employed with condenser-input filter systems, the voltage drop across the tube is constant at approximately 15 volts for all currents. This can be taken into account by calculating the performance as though the tube had zero resistance, and then reducing the direct-current voltage by 15 volts while keeping the load current, ripple voltage, and peak space current unchanged. The equivalent circuit of Fig. 273b also applies to inductance-input systems under conditions where the input inductance is insufficient to satisfy Eq. (198b), since the input inductance is then equivalent to added leakage inductance. This makes it possible to determine the amount of inductance that must be placed in series with the input lead to keep the peak space current

---

which the instantaneous current as given by the equations becomes zero, after which the current remains zero until the second anode becomes positive. It is to be noted that the current through the rectifier tube can never reverse in polarity even in the oscillatory case.

The d-c load current I_0 is the average current flowing through the rectifier, and so is obtained by averaging the pulse of current over a half cycle of the supply frequency. That is

$$I_0 = 2f\int_{t=0}^{t=t_1} i\,dt \tag{202a}$$

where i is the current given by Eq. (201), f is the supply frequency, and t_1 is the time at which the anode current becomes zero (see Fig. 273).

It will be noted that, in order to obtain a solution of Eqs. (200), it is necessary to know I_0, t_1, and t_0, and that at the same time I_0 depends upon i, t_1, and t_0. Because of the complexity of the relations involved, a straightforward solution of Eqs. (200) is impractical. The most satisfactory procedure for finding the current pulse corresponding to a load current I_0 consists in assigning a reasonable value to t_0 in Eq. (200a) or (200b) and then calculating and plotting the current from $t = 0$ up to the instant $t = t_1$ at which the current pulse becomes zero. The area under this curve of current is determined by a planimeter, and the average value over the length of time $1/2f$ is determined. This average represents the direct-current value of the pulse, and, if t_0 was correctly chosen, will equal I_0. If, as is usually the case, this is not true in the first trial, then further values of t_0 are tried until the correct value is found.

The voltage developed across the input condenser is:

$$\left.\begin{array}{l}\text{Voltage across } C\\ \text{from } t = 0 \text{ to } t = t_1\end{array}\right\} = E\sin\omega t_0 + \frac{1}{C}\int_0^{t} i\,dt - \frac{I_0 t}{C} \tag{202b}$$

$$\left.\begin{array}{l}\text{Voltage across } C\\ \text{from } t = t_1 \text{ to } t = \dfrac{1}{2f}\end{array}\right\} = E\sin\omega t_0 + \frac{I_0}{C}\left(\frac{1}{2f} - t\right) \tag{202c}$$

where i is given by Eq. (201a) or (201b), as the case may be, and the notation is as

down to a safe value when hot-cathode mercury-vapor tubes are employed with a condenser-input filter system.

The use of the curves of Fig. 274 in calculating the performance of condenser-input filter systems can be understood by the following example:

Example.—A load of 4000 ohms is to be supplied by direct current from a condenser-input rectifier filter system. The power transformer develops a potential of 300 volts r.m.s. on each half of the secondary. The effective leakage inductance and resistance are 40 mh and 125 ohms, respectively, and a Type 80 rectifier tube is used. The input condenser is 4 mf, and is followed by a single-section filter consisting of a 10-h choke with 400 ohms resistance and an 8-mf condenser across the output.

The first step in the solution is the determination of the effective load and source resistances. The former is the actual load resistance plus the choke resistance, or $4000 + 400 = 4400$ ohms. The average resistance of the rectifier tube is approximately $62/0.125 = 498$ ohms, so that the source resistance is $125 + 498 = 623$ ohms. Reference to Fig. 274 shows that, for these values of source and load resistances and input capacity, one has

	For $L = 0$	For $L = 0.1$	For $L = 0.040$
E_0/E	0.73	0.73	0.73
I_{max}/I_0	2.9	3.00	2.95
E_R/E_0	0.12	0.115	0.127

The last column is obtained by interpolation between the first two. From this column one has:

D-c voltage across input condenser $= 0.73 \times 300\sqrt{2} = 309$
D-c voltage across actual load $= 309 \times 4000/4400 = 281$
D-c current in load $= 309/4400 = 0.0703$ amp.
Peak space current $= 0.0703 \times 2.95 = 0.207$ amp.
Ripple voltage across input condenser $= 0.127 \times 309 = 39$ volts crest
Ripple voltage across load [from Eq. (197*b*)] $= \dfrac{39}{(754)^2 \times 10 \times 8 \times 10^{-6}} =$ 0.86 volts crest

above. The direct-current voltage E_0 developed across the input condenser is the average of the condenser voltage, or

$$E_0 = E \sin \omega t_0 + \frac{I_0}{4fC} - \frac{I_0 t_1}{C} + \frac{2f}{C}\int_0^{t_1}\int_0^{t} i \, dt \, dt \tag{202d}$$

The ripple voltage across the input condenser can be obtained by harmonic analysis of the wave represented by Eqs. (202*b*) and (202*c*).

The principal approximations involved in the above analysis are: (1) the rectifier is replaced by a fixed resistance during its conducting period; (2) the current flowing into the first inductance is assumed to be constant throughout the cycle; (3) the source of power exciting the primary of the power transformer is assumed to have zero internal impedance. The errors that result from these approximations are small under practical circumstances, and the method of analysis is capable of giving an accuracy that is sufficient for all ordinary design purposes. The use of a constant

Filters with shunt-condenser inputs are generally employed in radio receivers, small public-address systems, etc., when the amount of direct-current power required is small. Compared with inductance-input filters, the condenser-input arrangement produces a higher direct-current voltage, but has poorer regulation, higher peak space currents, and poorer transformer utilization. As a consequence the condenser-input arrangement is not favored where large power is required or where regulation of the direct-current voltage is important.

99. Rectifier-filter Systems. Miscellaneous Comments.—The rectifier and filter connections that should be used in a particular case depend upon the individual circumstances. Polyphase rectifiers are preferred when the direct-current power is in the order of several kilowatts or more because the ripple in the output of a polyphase rectifier is of small magnitude and of high frequency and so can easily be filtered out. Smaller outputs are most conveniently obtained from single-phase full-wave rectifiers, usually of the center-tapped transformer type shown in Fig. 266*b*.

Filters with series-inductance inputs are always used in polyphase circuits and are also preferred with high-power single-phase rectifiers because of the high ratio of average to peak current that is obtained with a series-inductance input. The use of a shunt-condenser input causes the ripple voltage appearing across the filter input to be less than with the series-inductance input, but the ratio of average to peak rectifier current is smaller, so that the condenser input is employed only in low-power single-phase rectifiers, such as those used to supply plate power for broadcast receivers. The filter inductances may ordinarily be placed in either the positive or negative lead.[1]

The condensers used in filters must be capable of *continuously* withstanding a direct-current voltage that is equal to the crest alternating-current voltage applied to the rectifier. Ordinarily a single condenser should be used to withstand the entire voltage, rather than several condensers in series. When condensers are in series, the direct-current voltage stress divides between them in proportion to their leakage resist-

---

resistance to represent the voltage drop in the rectifier tube is perhaps the most serious of the approximations, but, if the resistance is chosen so that the voltage drop produced by the peak rectifier current is equal to the voltage drop actually across the tube at this instant, the error will be small.

This footnote is based largely on material in the thesis submitted by Dr. R. L. Freeman in partial fulfillment for the doctor's degree at Stanford University.

[1] However, if it is essential that the hum output be extremely small, then the filter chokes must be placed in the ungrounded lead or there will be a residual hum caused by the capacity to ground of the transformer secondary winding. See F. E. Terman and S. B. Pickles, Note on a Cause of Residual Hum in Rectifier-filter Systems, *Proc. I.R.E.*, 22, p. 1040, August, 1934.

ances rather than their dielectric strength, and the leakage resistances are variable and uncertain.

The inductance coils used in the filter must have laminated iron cores, with an air gap that is sufficient to prevent the direct current from saturating the iron. The inductance that is effective to the alternating currents will vary with the superimposed direct current as discussed in Sec. 6 and will normally be greatly lowered by the presence of the direct current. The insulation between the winding and the core must be capable of withstanding the full direct-current voltage delivered by the rectifier output.

The transformers used in rectifier-filter systems that supply power to radio receivers and high-gain audio-frequency amplifiers are often provided with grounded electrostatic shields between the primary and secondary. Such a shield prevents disturbances that originate on the power line from being transferred to the output through capacity between the primary and secondary windings.

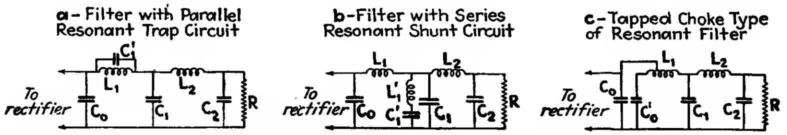


FIG. 276.—Filter circuits using tuned elements.

The effectiveness of a filter in suppressing a particular frequency can be improved by tuning one of the series inductances to parallel resonance at the desired frequency as shown in Fig. 276*a*, or by tuning one of the shunt capacities to series resonance as shown in Fig. 276*b*. Such arrangements, however, have the disadvantage that the tuning reduces the attenuation of the other frequency components of the ripple, and also that each resonant circuit must be individually adjusted.

The tapped-choke arrangement of Fig. 276*c* was at one time extensively used in radio receivers; it has the advantage of giving less hum with the same total filter condenser capacity than do conventional circuits. For effective hum suppression the capacity C_0' must be so adjusted in relation to the tap on the coil that the branch through C_0' offers a capacitive reactance such that the alternating-current hum voltage developed across the left-hand section of the coil will induce a voltage in the right-hand portion of the coil which exactly bucks out the hum voltage applied to the filter. Under such conditions there is no a-c current flowing in the right-hand side of the coil and there is a complete cancellation of the hum having the frequency for which the adjustment has been made. While the tapped-choke arrangement is both simple and effective, the introduction of electrolytic condensers has

shifted the economic balance in such a way that this system is now seldom used.

Filters using series resistance instead of series inductance are commonly employed where high voltages and low currents are involved, or where it is necessary to drop the voltage. Such an arrangement is illustrated in Fig. 277 and results in a suppression given approximately by the formula

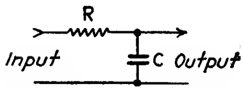


FIG. 277.—Resistance-condenser filter.

$$\frac{\text{Alternating-current voltage arcoss load}}{\text{Alternating-current voltage applied to input}} = \frac{1}{R\omega C} \tag{203}$$

where

R = series resistance

C = shunting capacity following the series resistance

ω = 2π times the ripple frequency involved.

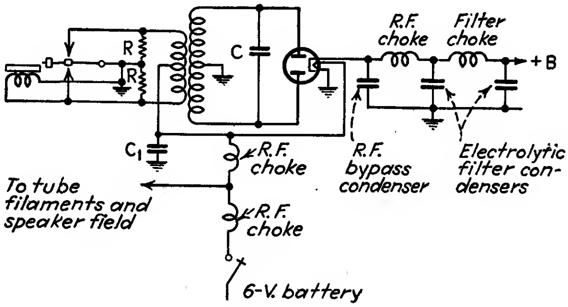


FIG. 278.—Circuit diagram of non-synchronous vibrator power supply.

100. Miscellaneous Supply Systems. *Vibrator Systems.*[1]—The vibrator power-supply system produces direct-current plate power from a storage battery by using a vibrating contact to change the direct current from the battery into alternating current which can be stepped up in voltage by a transformer and rectified to produce high-voltage direct-current power. A typical vibrator circuit is shown in Fig. 278, in which the vibrator operates on the same principle as the ordinary buzzer. The resistances R and the condenser C_1 are for the purpose of preventing sparking at the contacts, while the condenser C absorbs transients that would otherwise produce radio interference, and the radio-frequency chokes prevent radio-frequency surges from getting into the battery.

Another form of vibrator power system known as the synchronous type is shown in Fig. 279. Here the vibrating armature has a pair of

[1] For further information see Walter Van B. Roberts, Vibrator Power Supply from Dry Cells, *Electronics*, vol. 7, p. 214, July, 1934; Vibrators, *Electronics*, vol. 9, p. 25, February, 1936.

contacts that rectify the output voltage of the secondary, thereby dispensing with the rectifier tube required in the non-synchronous arrangement of Fig. 278.

The vibrator represents an inexpensive means of obtaining plate power from a storage battery, but has the disadvantage of a limited life and of requiring a relatively complicated filtering and shielding system to prevent surges from producing interference in the radio receiver. The synchronous type of vibrator is more complicated than the non-synchronous type, but eliminates the rectifier tube with its voltage drop and so is used about as frequently as the non-synchronous arrangement.

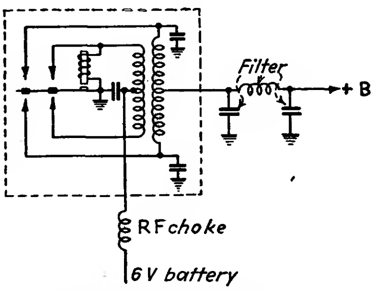


FIG. 279.—Circuit diagram of synchronous vibrator power supply.

Dynamotor.—A dynamotor is a combined motor generator having two or more separate armature windings and a common field circuit. One of the armature windings is operated from a low-voltage direct-current power source, commonly a storage battery, to give motor operation, while the other windings then serve as generators to produce the desired direct-current voltages.

Dynamotors using permanent magnetic fields and operating from a storage battery are often used for supplying the anode power required by automobile radio receivers. Compared with the vibrator systems the dynamotor has the advantage of longer life and less trouble from radio-frequency surges, but is considerably more expensive.

Self-rectifying Circuits.—If an alternating voltage is applied directly to the plate of an oscillator in place of the usual direct-current potential, the oscillator will operate on the positive half cycles of the applied voltage and so will generate wave trains having the character shown in Fig. 280. These are known as interrupted continuous waves (abbreviated I.C.W.) and are occasionally used in radio telegraphy. Another type of self-rectifying oscillator circuit is shown in Fig. 281, in which the two plates of a two-tube oscillator are supplied with alternating-current voltages that are 180° out of phase, and a large inductance is inserted in the common filament return lead. This arrangement is electrically equivalent to a single-phase full-wave rectifier circuit with a filter consisting of a single inductance (see Fig. 268*a*). When the inductance is large, the total current drawn by the two tubes is substantially constant, and the generated oscillations will also be of constant amplitude. Practically, however, it is impossible to maintain the total current entirely constant, and as a result modulated oscillations are generated.

Self-rectifying oscillator circuits find their chief usefulness under conditions where the complications of a rectifier-filter system are to be avoided. These circuits are not suitable for ordinary commercial radio transmitters because the voltage that is applied to the plates of the oscillators is not constant, and as a result the generated oscillations suffer both frequency and amplitude modulation.

Anode Power for Aircraft Radio Equipment.—Aircraft radio receivers commonly obtain plate power from a small dynamotor operating from the 12-volt storage battery on the plane. Anode power for airplane

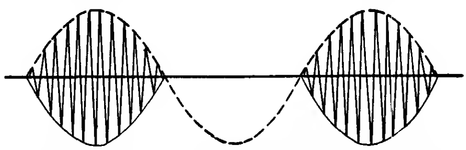


FIG. 280.—Interrupted continuous waves generated by applying an alternating-current voltage directly to the plate of an oscillator.

transmitters is more of a problem, however, because the power requirement is large. The most common arrangement employs a dynamotor operating from the 12-volt storage battery of the plane, together with a battery-charging generator operated from the airplane engine. This arrangement has the disadvantage of giving only about 30 min. emergency service when the battery is fully charged but with the plane engines not operating. In large planes the demand for electrical power for lighting, starters, fuel pumps, etc., is becoming so heavy that the present trend is toward the use of a generator driven by an auxiliary gasoline engine.

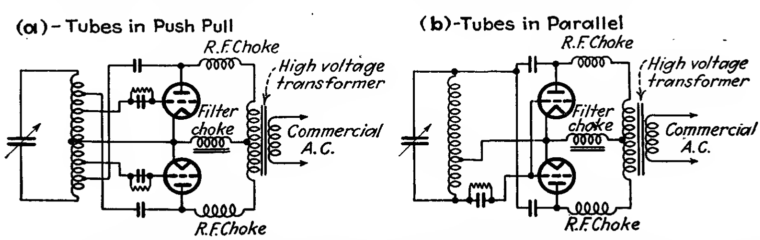


FIG. 281.—Self-rectifying oscillator circuits capable of generating oscillations of substantially constant amplitude if the filter choke is very large.

Problems

1. Explain the alternating-current hum effects that are produced in a filament-type radio-frequency amplifier tube that is operated from alternating current, keeping in mind that the radio-frequency coupling circuits will not transfer audio-frequency voltages.

2. Determine a suitable by-pass condenser for the bias resistor of a resistance-coupled amplifier using a 6C6 tube and designed according to the fourth column of Table VII, Chap. V, assuming that the regeneration from the bias impedance is not to affect the amplification seriously for frequencies above 60 cycles.

3. In the resistance-coupled amplifier of Fig. 97 the bias resistor is 6000 ohms. Calculate the actual amplification curves for $C_c = 0.01$ mf when the by-pass condenser in parallel with the bias resistance is (*a*) 1.0 mf, (*b*) omitted.

4. Discuss the practicability of using a self-bias arrangement consisting of a resistor between cathode and ground for (*a*) a Class C amplifier, (*b*) a linear amplifier, and (*c*) a Class B audio amplifier.

5. In a push-pull Class A audio amplifier, explain why it is that no by-pass condenser is needed in shunt with the bias resistor between cathodes and ground.

6. *a*. Explain why in a Class AB push-pull amplifier the bias arrangement of Fig. 259*c* is equivalent to a self-bias arrangement such as Fig. 259*a* and *b*, while the arrangement at *d* behaves more nearly as a battery bias.

b. On the assumption that the anode voltage is obtained from a rectifier-filter system, discuss the relative power-transformer secondary voltages required for the various bias arrangements of Fig. 259.

7. Explain why it is customary to design a rectifier tube with a maximum permissible plate dissipation that is less than the plate dissipation that would result if the allowable peak current were drawn continuously through the tube by applying a direct-current anode voltage.

8. If the negative lead of a power-supply system is to be grounded, discuss the difference in insulation requirements of the transformer secondary winding for (*a*) the single-phase circuits of Fig. 266, (*b*) the polyphase circuits of Fig. 267.

9. Describe in detail the way in which the three-phase full-wave rectifier circuit of Fig. 267*c* operates.

10. From the information given in Table XIII and Fig. 266 discuss the relative merits of the full-wave bridge *versus* the full-wave center-tap circuit for single-phase rectification where an inductance input is used in the filter.

11. *a*. Check the results given in Table XIII for the single-phase full-wave center-tapped connection.

b. Check the results given in Table XIII for the three-phase half-wave circuit.

12. The power supply for the final power amplifier of the transmitter of Fig. 282 employs six Type 869A hot-cathode mercury-vapor tubes in a three-phase full-wave circuit operating from 60 cycles. The power supply is rated at 18,000 volts and 4 amp., and has a single-section inductance-input filter consisting of 0.5 henries and 3.0 mf. Neglecting the resistance of the filter inductance and the leakage reactance and resistance of the power transformer, calculate (*a*) peak current through the rectifier tubes, (*b*) peak inverse voltage, (*c*) secondary voltage of power transformer, (*d*) kva rating of transformer primary and secondary windings, (*e*) amount of 360-cycle and 720-cycle ripple voltage appearing in the output.

13. *a*. In the power-supply system of Prob. 12, determine total power loss in the six rectifier tubes, including both filament power and plate dissipation, and from this calculate the rectifier efficiency. In evaluating the tube loss assume a tube drop of 12 volts.

b. Assume the Type 869A hot-cathode mercury-vapor tubes of (*a*) are replaced by Type 214 high-vacuum tubes, which have a tube drop of approximately 1000 volts for the current which will flow through the tube with this power-supply system. Determine the total rectifier tube power (both filament power and plate dissipation). Calculate the over-all efficiency of the rectifier and the extra power required with high-vacuum tubes as compared with hot-cathode mercury-vapor tubes.

14. Design a power-supply system for operating a push-pull Class C amplifier using 852 tubes (allowable plate loss 100 watts per tube), and operating at $E_b =$ 3000 volts. The maximum allowable ripple voltage is 2 per cent and the available power is 60 cycle, single phase. Use hot-cathode mercury-vapor rectifier tubes.

15. Two Type 46 tubes are to be used in a push-pull circuit as Class B audio amplifiers, with a plate potential of 400 volts. The plate current drawn by the combination with no applied signal is 12 ma, and with full signal is 130 ma. Design a suitable inductance-input filter system which will maintain good regulation from no signal to full signal and which will not have more than 1 per cent ripple voltage.

16. Explain in a qualitative manner why in a condenser-input filter (*a*) the direct-current output voltage increases with load resistance, (*b*) the ripple voltage across the input condenser decreases with increased load resistance, and (*c*) the ripple voltage becomes more and the direct-current voltage less when a smaller input condenser is used?

17. *a.* In a particular condenser-input rectifier-filter system the power transformer develops a potential of 375 r.m.s. volts across half the secondary, while the effective transformer resistance and leakage inductance referred to one-half of the secondary are respectively 92 ohms and 28 mh. If a 4-mf input condenser is employed with a 5Z3 tube, calculate and plot direct-current voltage and ripple voltage across the input condenser, and the peak anode current as a function of load resistance.

b. Design a filter system to follow the input condenser which will make the ripple in the filter output less than 0.01 volt when the load resistance is 2500 ohms.

18. Explain why electrolytic condensers are extensively used for filter condensers up to about 500 volts peak and for cathode by-pass condensers, but are not suitable for blocking condensers, etc.

19. In a vibrator supply system such as shown in Fig. 278, sketch oscillograms showing (*a*) current flowing in each half of primary, (*b*) voltage across each half of the primary, and (*c*) voltage across the secondary winding.

20. Describe the operation of the self-rectifying circuit of Fig. 281, with particular reference to the action of the choke in the common lead. In this discussion sketch oscillograms illustrating the general character of the plate voltage and plate current in each tube and the voltage and current in the choke.

CHAPTER XII

RADIO TRANSMITTERS

101. Radio Transmitters. General Considerations.—A radio transmitter is essentially a device for producing radio-frequency energy that is controlled by the intelligence to be transmitted. A transmitter accordingly represents a combination of oscillator, power amplifiers, harmonic generators, modulator, power-supply systems, etc., which will best achieve the desired result. Inasmuch as these various elements which go to make the completed transmitter have already been individually considered, the present chapter is devoted to a consideration of combinations most commonly employed in commercial equipment.

Commercial transmitting equipment is ordinarily mounted on a framework of structural-steel members fronted by a vertical metal panel containing the controls and meters necessary for adjusting and monitoring the transmitter. All equipment appearing on the panel is at ground potential, and instruments which must be observed during adjustment or operation and which are not at ground potential are located behind the panel and viewed through windows. The steel frame is normally enclosed with wire mesh of some sort and is provided with doors that cut off the transmitter power when opened. Typical examples of this general type of construction are to be found in Figs. 283, 284, and 289. This type of construction requires a minimum of floor space in proportion to the amount of apparatus involved, makes the transmitter accessible for inspection and repairing, and eliminates all hazard to persons.

The design of most radio transmitters, particularly those intended for broadcast and short-wave transmission, is dominated by the need of maintaining the transmitted frequency as nearly constant as possible over long periods of time. In broadcast work two or more transmitters are commonly assigned the same carrier frequency, and in order to minimize the resulting interference it is essential that the carrier frequencies be as nearly as possible the same. In short-wave work, particularly with code transmitters, the side bands required are so narrow that the frequency separation between transmitting stations is determined primarily by the frequency stability of the carriers. For example, a code transmitter operating at a speed of 1000 letters per minute requires, according to Sec. 72, a side band 131 cycles wide. With perfect stability

of the carrier frequency two such transmitters could operate on frequencies differing by only 300 cycles, and a relatively limited space in the frequency spectrum could accommodate a large number of transmitters. Actually, however, it is impracticable under commercial conditions to maintain the carrier frequency of a short-wave transmitter closer than about 0.03 per cent to the assigned value, so that, if the carrier frequency is 20,000 kc, the actual frequency spacing must be of the order of 12,000 cycles. The importance of frequency stability is more or less proportional to the carrier frequency, since a given percentage of variation represents a smaller number of cycles at a low than at a high carrier frequency.

102. Code Transmitters.—The design of a code transmitter depends primarily on the frequency stability required and the power to be generated. Where the frequency stability must be high, as is the case with most types of short-wave code transmitters, the entire design must be built around some form of oscillator that will maintain a constant frequency, such as a crystal oscillator, a resonant line oscillator, etc. Where the frequency stability need not be particularly great, as is the case in long-wave transmitters and in certain classes of short-wave equipment such as amateur transmitters, it is often possible to introduce a number of simplifications by sacrificing frequency stability. Typical examples of the most commonly used types of code transmitters are discussed in the following paragraphs.

Crystal-controlled Commercial Short-wave Code Transmitters.—When the frequency stability of the transmitted carrier is of importance, one of the most satisfactory arrangements is to employ a crystal oscillator to control the frequency. With such an arrangement the entire design must be built around the fact that the amount of power that can be controlled by a crystal is relatively low, while the highest frequency obtainable commercially from a crystal is of the order of 6000 kc. Greater power requires the use of Class C amplifiers, while higher frequencies involve the use of harmonic generators. The usual crystal-controlled short-wave transmitter makes use of a crystal oscillator involving a tube having a rating of the order of 7½ watts and generating a frequency in the range 1500 to 4000 kc. The crystal is followed by an amplifier tube, commonly termed a buffer amplifier, which isolates the crystal from the remainder of the transmitter. In order to be sure of the highest possible frequency stability, the buffer tube is usually operated so that it draws little or no grid current, and the crystal and buffer tubes normally have a separate power supply and operate continuously even when the key of the transmitter is up. The buffer tube is followed by harmonic generators and Class C amplifiers as required to develop the required power output and frequency.

The characteristic features of a well-designed crystal-controlled short-wave code transmitter are illustrated by the high-power transmitter shown in Fig. 282. This particular transmitter is designed to operate in the frequency range 3000 to 21,500 kc (100 to 14 meters wave length) and is capable of delivering a power output of 30 to 40 kw. The transmitter itself is divided into two parts, the first of which is an exciter unit containing the crystal oscillator, buffer amplifier, harmonic generators, and power amplifiers capable of delivering 1-kw output. The second unit is a one-stage power amplifier using water-cooled tubes, and it is connected to the exciter unit by a short transmission line. This arrangement, with the final power amplifier separated from the remainder of the transmitter, is frequently used in high-power equipment because it simplifies the problem of shielding the low-power-level circuits from the strong fields produced by the final power amplifier. The exciter unit can be considered as a complete transmitter of moderate power, and its output can be delivered directly to an antenna if desired.

The sequence of tubes in the exciter unit is as follows:

7½-watt triode crystal oscillator.
75-watt screen-grid buffer amplifier.
75-watt screen-grid frequency doubler.
75-watt screen-grid frequency doubler (second doubler).
500-watt screen-grid tube used as frequency doubler when output frequency is in excess of 12,000 kc, and otherwise employed as power amplifier.
Two 500-watt screen-grid tubes in push-pull amplifier circuit.

The final power amplifier unit consists of two water-cooled tubes in a neutralized push-pull circuit.

The crystal oscillator employs the circuit shown in Fig. 206, and is followed by a buffer amplifier consisting of a screen-grid power tube. By the use of suitable shielding and a separate power supply for the crystal oscillator and buffer tubes, the generated frequency is virtually independent of the conditions existing in the remainder of the transmitter. In order to prevent the frequency from varying with temperature in the case of X- or Y-cut crystal plates, an oven is provided for maintaining the crystal temperature constant, and the crystal is operated at a relatively low amplitude so the internal heating due to its own losses will be small.

The harmonic generators and amplifiers following the buffer tube are of the Class C type and increase the crystal output to approximately 1 kw. For frequencies above 12,000 kc the three tubes following the buffer amplifier are all adjusted to generate second harmonics, and the crystal is ground to generate a frequency one-eighth of that to be transmitted. For frequencies in the range 6000 to 12,000 kc the crystal is ground to one-fourth of the transmitted frequency, and only two harmonic genera-

tors are employed, the third tube being used as an ordinary Class C amplifier. Likewise in the frequency range 3000 to 6000 kc only one

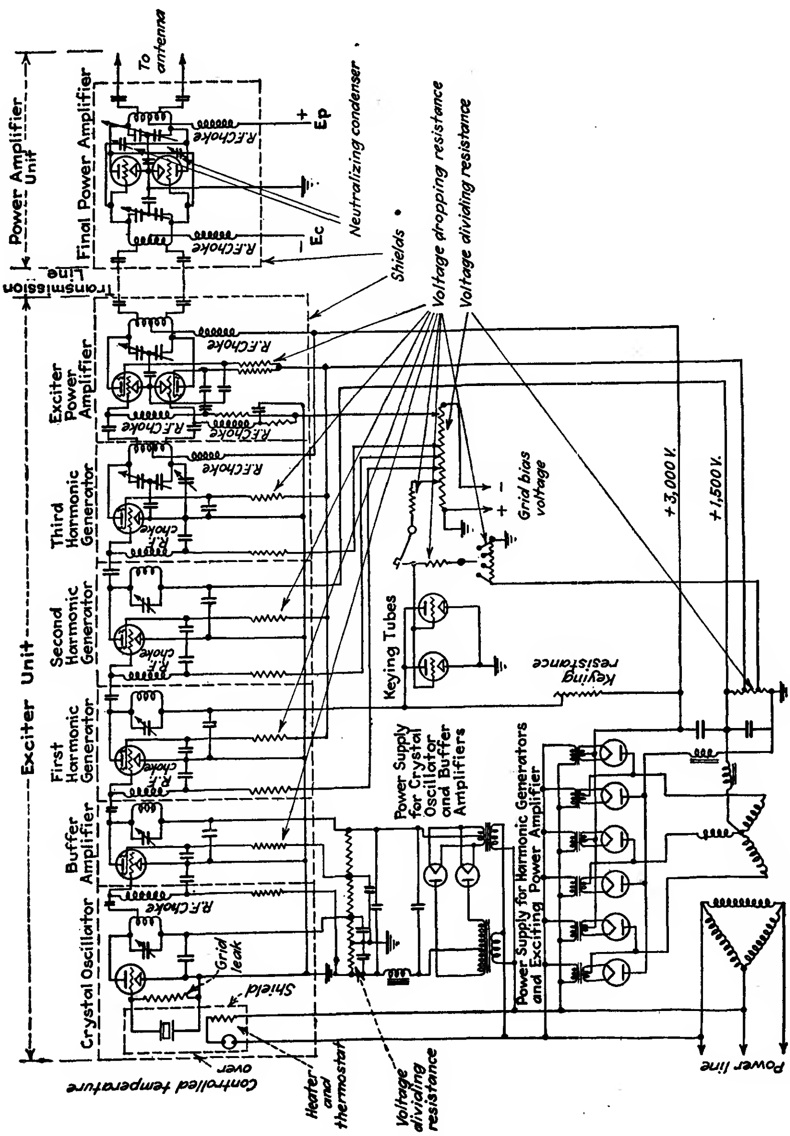


FIG. 282.—Simplified circuit diagram of a representative high-power, short-wave code transmitter (RCA 20–40 kw transmitter).

harmonic generator is employed, and the remaining tubes are used as Class C amplifiers.

Tuning is accomplished in the exciter unit by means of variable condensers, changing inductance coils as required, and by short-circuiting

turns of the inductance coils. In the power-amplifier unit rough tuning can be accomplished by changing coils and by adjusting the tuning capacity in steps, while fine tuning makes use of a copper ring rotated in the field of the inductance. Several crystals are provided in the crystal holder, and switches operated from the panel make it possible to have available two frequencies of transmission, with the change over from one to the other requiring only a few seconds.

Three separate anode power-supply systems are provided. The crystal oscillator and buffer amplifier operate from a single-phase center-

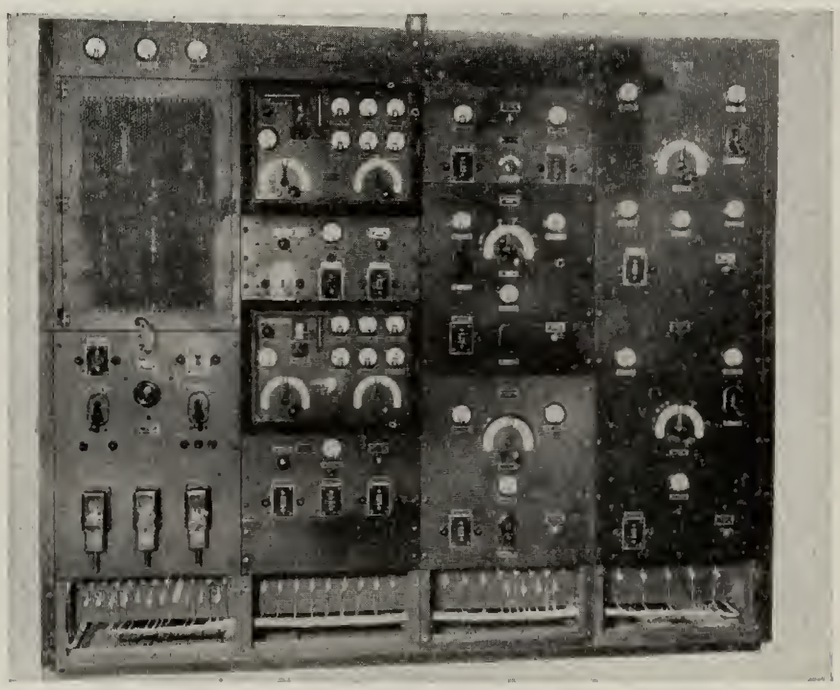


FIG. 283a.—Exciter unit of transmitter of Fig. 282—front view of panel.

tapped rectifier using Type 866 mercury-vapor tubes. A voltage dividing resistance across the output of the rectifier makes it possible to obtain anode, screen-grid, and control-grid bias voltages from a single rectifier. The remainder of the exciter receives its anode power from a three-phase full-wave rectifier using Type 872 mercury-vapor tubes in conjunction with a filter consisting of a series input inductance followed by a shunt condenser. This unit delivers 3000 volts direct current, which is applied directly to the plates of the 500-watt screen-grid power-amplifier tubes. Lower voltages for the other tubes and for the screen grids are obtained either by means of voltage-dropping resistances from the 3000-volt source or from a 1500-volt tap obtained by connecting to

the neutral point of the secondary star through a filter inductance. This tap causes three of the six rectifier tubes to do double duty by also functioning in a three-phase half-wave circuit to give 1500 volts output. The grid-bias voltages required by the exciter are obtained either from a 1000-volt motor generator or from a suitable rectifier-filter arrangement with the aid of a voltage-dividing resistance. The power-amplifier unit has a separate three-phase full-wave rectifier using Type 869A mercury-

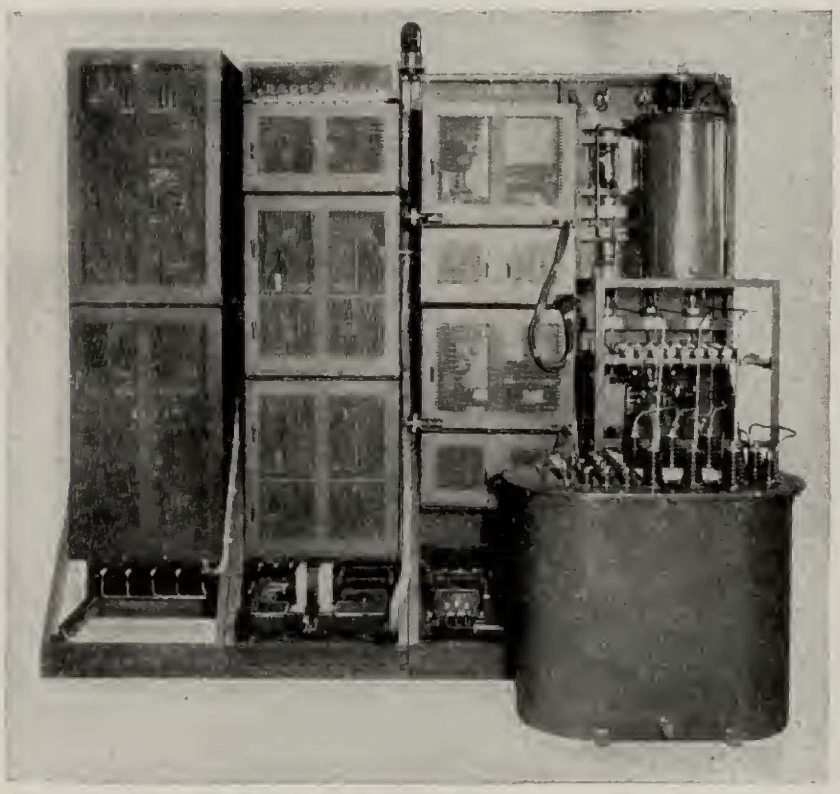


FIG. 283*b*.—Exciter unit of transmitter of Fig. 282—rear view with shields in place.

vapor tubes and delivering a direct-current output of 4 amp. at 18,000 volts. The filter consists of a 0.5-henry series input inductance followed by a 3-μf shunt condenser.

It is to be understood that this description of the transmitter has dealt only with the outstanding features and has omitted innumerable details that must be taken into account in the actual construction in order to obtain satisfactory results.[1] For example it is necessary to

[1] For further details concerning an earlier but only slightly different version of this transmitter, see I. F. Byrnes and J. B. Coleman, 20–40 Kilowatt High-frequency Transmitter, *Proc. I.R.E.*, vol. 18, p. 422, March, 1930.

shield the parts of the transmitter from each other by means of copper shields, the location of which is indicated by the dotted lines of Fig. 282. Radio-frequency choke coils and by-pass condensers must also be provided in many places not shown in this schematic diagram if parasitic oscillations and regeneration are to be avoided, and it is also necessary to arrange the various parts of the transmitter in relation to each other in such a way as to minimize undesirable stray couplings while making the connections as short and as direct as possible.

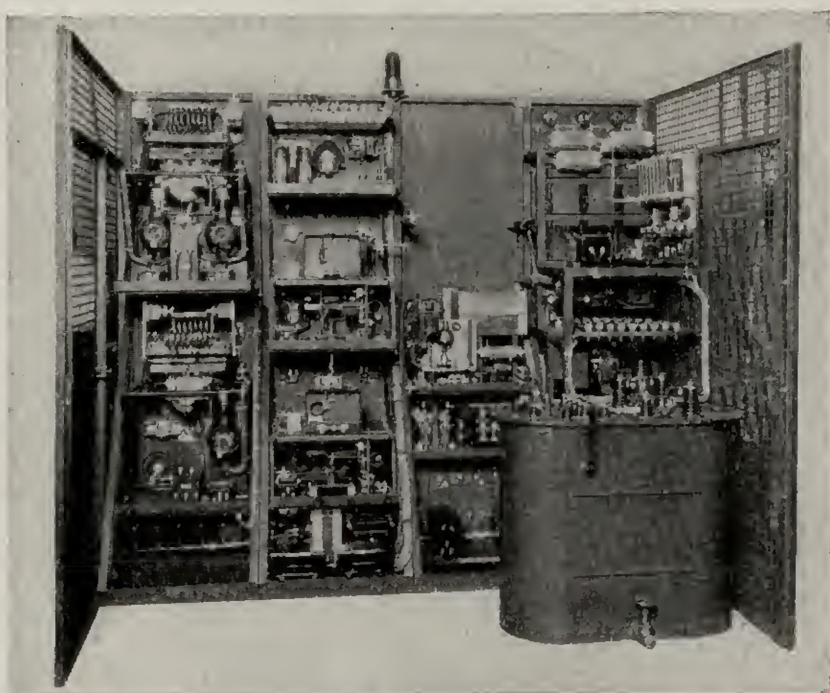


FIG. 283c.—Exciter unit of transmitter of Fig. 282—rear view with covers of shields removed.

Photographs showing different views of the complete transmitter are given in Figs. 283 and 284 and serve to bring out the more obvious features of construction.

While the high-power transmitter that has been described incorporates all the typical features of a short-wave crystal-controlled code transmitter, it is of course possible to modify the layout in many details. Thus the buffer tube is often made to serve as a harmonic generator by tuning its tank circuit to twice the frequency generated by the crystal. Triodes are also commonly used as harmonic generators instead of screen-grid tubes, and can be used as radio-frequency amplifiers in place of screen-grid tubes provided neutralized circuits are employed. In

some cases the harmonic-generator tubes are called upon to generate the third or even the fourth harmonic instead of the second, although the power output on these higher harmonics is small. Sometimes tuned circuits use only the capacities supplied by tubes, leads, etc., and are adjusted to resonance by varying the circuit inductance.

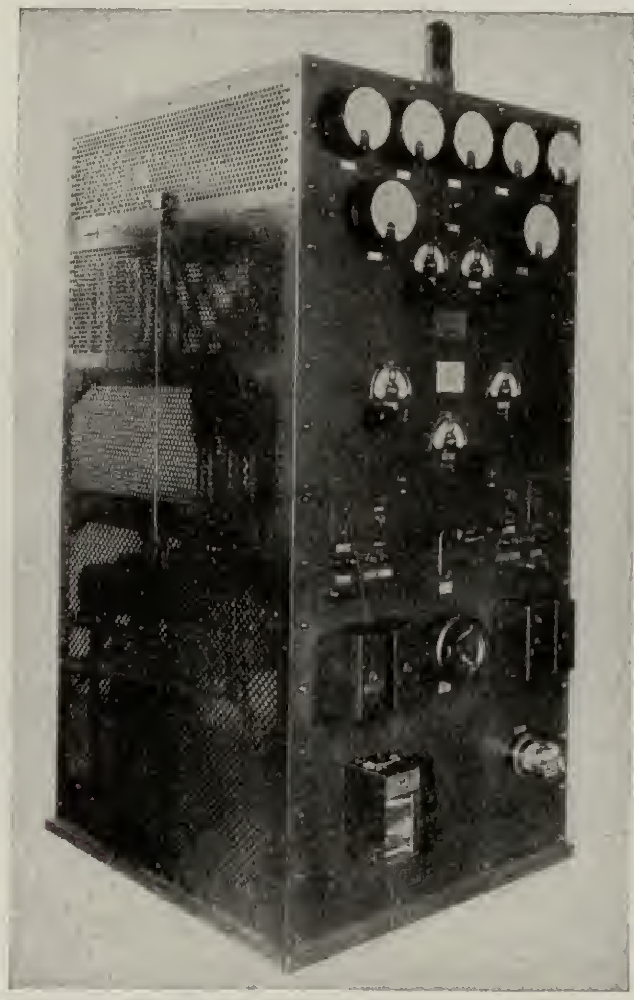


FIG. 284a.—Front view of 30- to 40-kw output power amplifier.

Crystal-controlled Short-wave Amateur Code Transmitter.—Amateurs interested in short-wave telegraphy have developed a great variety of low-power code transmitters. The better grade of these employ crystal oscillators followed by harmonic generators and Class C amplifiers as required to develop the desired frequency and output power. A typical

transmitter of this class is illustrated in Fig. 285, and makes use of a crystal oscillator followed by a harmonic generator and a push-pull Class C amplifier using neutralized triodes. Descriptions of innumerable transmitters of this type are to be found in the amateur's handbooks and magazines, and many of these transmitters represent great ingenuity

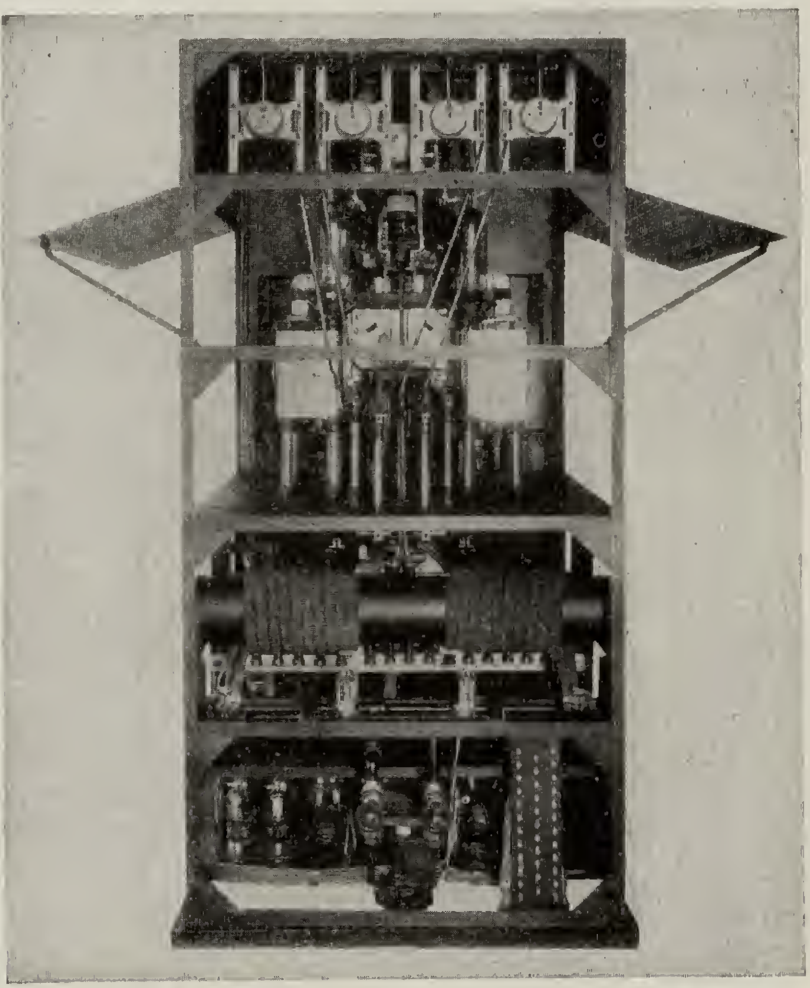


FIG. 284*b*.—Rear view of 30- to 40-kw output power amplifier.

in adapting available tubes to give a maximum power output in proportion to cost. While amateur equipment is usually designed to operate less conservatively than commercial equipment, the essential features are the same as for commercial equipment; in fact many low-power commercial transmitters are essentially commercial adaptations of amateur designs.

Master-oscillator Power-amplifier Transmitters.—Master-oscillator power-amplifier arrangements are commonly used in transmitters operating at frequencies below 500 kc, since the percentage frequency stability required at these lower frequencies is less than with short-wave trans-

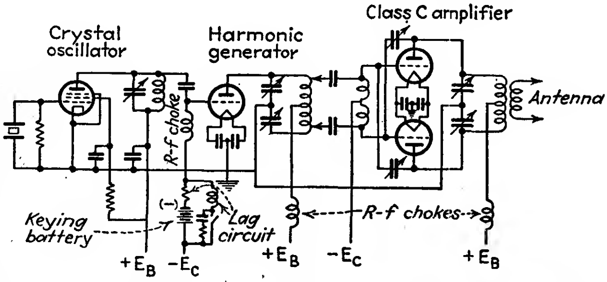


FIG. 285.—Typical amateur crystal-controlled code transmitter.

mitters. Master-oscillator power-amplifier transmitters are also employed to some extent in portable equipment, in amateur work, and in other circumstances where the transmitter is required to operate somewhere within a band of frequencies but is not required to maintain any exact frequency to a high degree of precision.

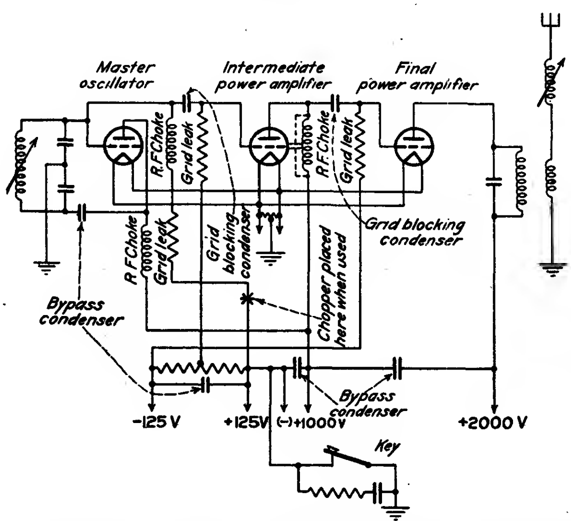


FIG. 286.—Simplified circuit diagram of 2-kw ship transmitter designed for use in the frequency range 125 to 500 kc. This transmitter consists of a master oscillator and intermediate and output amplifiers and is keyed by opening the ground connection of the plate and grid return leads.

Compared with crystal-controlled transmitters, the master-oscillator power-amplifier arrangement has the advantage of simplicity. The master-oscillator can operate at the transmitted frequency and can be designed to develop sufficient power to excite the final power amplifier

directly. This eliminates the chain of frequency multipliers and Class C amplifiers commonly required to develop large powers when crystal control is employed.

A typical master-oscillator power-amplifier transmitter for marine radio-telegraph communication at frequencies below 500 kc is shown in Fig. 286. This consists of a master oscillator, followed by a buffer power amplifier which excites a final Class C power amplifier. A typical low-power master-oscillator power-amplifier short-wave transmitter is illustrated in Fig. 287 and consists of a master oscillator which excites a screen-grid Class C power amplifier directly.

The frequency stability of a master-oscillator power-amplifier arrangement is usually less than that of a crystal-controlled transmitter because

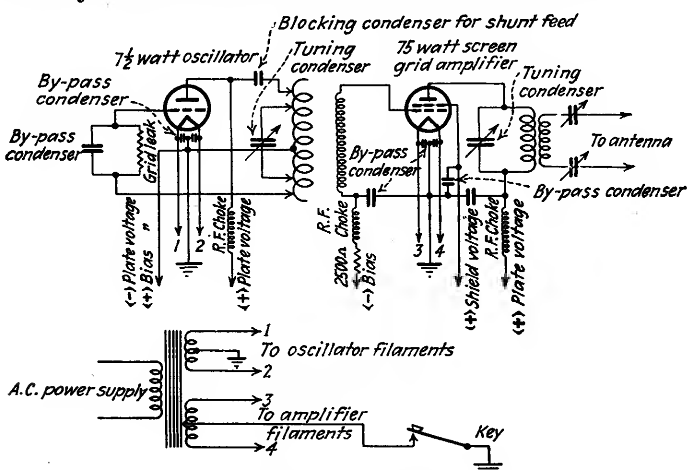


FIG. 287.—Circuit diagram of typical low-power master-oscillator power-amplifier short-wave transmitter.

of the variation in oscillator frequency with temperature, electrode voltages, etc., but, by careful attention to design details and by placing voltage regulators in the power-supply systems, the frequency stability can be made quite high, in fact, adequate to meet the governmental regulations for code transmitters. If, in addition, the master oscillator and its associated tuned circuits are placed in a compartment that is maintained at constant temperature, the frequency stability can be still further improved. Transmitters of this latter type, involving a low-power tuned-circuit oscillator followed by harmonic generators and Class C amplifiers, have a frequency stability comparable with that of the ordinary crystal-controlled transmitter.

Transmitters Employing a Simple Oscillator.—Simple oscillators delivering their output directly to the antenna have the disadvantage that any change in the load, such as might be produced by wind blowing

on the antenna, reacts upon the oscillator and tends to change the frequency. It is consequently necessary with such transmitters to arrange matters so that the inherent frequency stability is as high as possible and so that the load is essentially a resistance that is relatively loosely coupled to the tank circuit of the oscillator.

103. Keying of Code Transmitters.—The output of a high-power crystal-controlled transmitter is ordinarily turned on and off in accordance with the characters of the telegraph code by means of a keying system which operates on one of the low-level amplifier or harmonic generator tubes in such a way as to remove the alternating-current excitation from the grids of the large tubes. The full power output is thus controlled by keying only a small amount of energy and a low-power relay can be used. The crystal oscillator and its buffer tube normally have an independent source of anode power and are allowed to operate continuously in order that the generated frequency may reach an equilibrium value that is unaffected by the rapidity and character of the keying.

There are a number of ways by which a power-amplifier or harmonic-generator tube may be prevented from delivering output to a succeeding tube. A typical example is illustrated in the schematic diagram of the high-power transmitter shown in Fig. 282. Keying is accomplished here in the first doubler stage by reducing the plate voltage supplied to the doubler tube to the point where the output is insufficient to bring the instantaneous grid potential of the succeeding tube above cut-off. The keying unit consists of two 50-watt tubes in parallel, the plates of which are fed in parallel with the plate of the first doubler tube through a common resistance from the 3000-volt power supply. When the key is down, a negative bias exceeding the cut-off value is placed on the grids of the keying tubes so that the keying unit draws no current and allows normal voltage to be applied to the plate of the doubler tube. When the key is up, a slightly positive voltage is applied to the grids of the keying tubes, causing these tubes to draw a large plate current through the series resistance and hence to reduce the potential that is applied to the plate of the doubler tube to a very low value. Another method consists of an arrangement that places a large additional bias voltage on the grid of one of the amplifier or harmonic-generator tubes when the key is up and removes the extra bias when the key is down. The added bias is made sufficient to block the tube and so prevents operation of the transmitter.

In the master-oscillator power-amplifier type of transmitter the ideal keying arrangement would permit the oscillator to operate continuously from a separate power source and would key in the power-amplifier circuit. From a practical point of view, however, a separate power source for the master oscillator is seldom justified. A variety of methods

are employed for keying such transmitters, including the use of a relay in the high-voltage supply, a key or relay in the primary circuit of the plate-supply transformer, and various means for blocking one or more of the tubes. A typical arrangement of the latter type is illustrated in Fig. 286, and involves grounding the grid and plate return leads through the key.[1] When the key is down, the tubes operate in a normal manner; when the key is up, the plate and the filament assume the same potential (*i.e.*, ground potential), causing the negative plate lead and hence the grid potential to become negative with respect to the filament, thereby blocking the tubes. A modification of this method is shown in Fig. 287, where the key removes the ground from the amplifier filaments to block operation.

Transmitters in which there is an oscillator that delivers its power directly to the radiating system are usually keyed in the same manner as master-oscillator power-amplifier arrangements.

Keying Troubles.—A proper keying system gives clean-cut dots and dashes having constant carrier frequency and produces a minimum of interference. To achieve this it is necessary to pay proper attention to the details of the keying action.

Constant carrier frequency means freedom from frequency variations during the interval of a dot or dash, and also freedom from drift in frequency during operation. Frequency variation during a dot or dash produces an effect commonly termed a *chirp,* and may be the result either of power-supply systems with poor regulation or of rapid heating. Gradual drifts in frequency during transmission can be attributed to gradual heating of tuned circuits, tubes, etc., causing the frequency to drift steadily during use and to be different for continuous operation than after a period of rest. These slow drifts can be virtually eliminated in crystal-controlled oscillators by proper design, and in the case of other systems can be minimized by operating tubes and equipment conservatively and by sending out dots during otherwise idle intervals in order to maintain the temperature somewhere near normal.

Some keying systems fail to give clean-cut dots and dashes, particularly when the speed of keying is high. Thus, if the keying is done on the primary side of the plate power transformer, a brief interval must elapse after the key is depressed before the normal direct-current voltage is obtained from the filter output, while, after the key breaks the circuit, the transmitter will continue to operate for a brief time as a result of the energy stored in the filter inductances and capacities.[2] The result is

[1] For other keying methods in which one or more of the tubes are made inoperative, see "Radio Amateur's Handbook," American Radio Relay League.

[2] Other methods of keying also produce transients that distort the code characters somewhat. For example, in the keying arrangement of Fig. 282 the load on the

that the code characters are rounded off and tails tend to form after each impulse; with very fast keying there may be actual overlapping.

Code transmitters with poorly designed keying systems also commonly produce interference in near-by radio receivers even when these are tuned to an entirely different frequency from the transmitted carrier. This interference is in the form of thumps or clicks which occur as the key makes and breaks its circuit. Such key clicks are caused by high-order side-band frequencies produced by the sudden starting and stopping of oscillations. In a hypothetical case where the transmitted energy is assumed to be turned on and off instantly, the resulting wave can be considered as containing side-band components extending all the way from zero to infinite frequency, and, although the amount of energy at any one frequency is extremely small, interference is produced in near-by radio receivers. The remedy for key clicks is either to make the radiated energy pass through tuned circuits which suppress the high-order side-band components or to use a time delay or lag circuit in association with the keying so that the transmitted signals are turned on and off gradually rather than abruptly, thereby preventing the generation of high-order side bands. Either method is satisfactory. In the former it is necessary to provide proper shielding so that the undesired side-band components will not be radiated without passing through the tuned circuit; in the latter method it is necessary to proportion the lag circuit carefully to avoid seriously distorting the code characters, while at the same time suppressing the clicks.

104. Radio-telephone Transmitters.—In radio-telephone transmitters the output is modulated in accordance with the intelligence to be transmitted, instead of being simply turned on and off by means of a key. The usual radio-telephone transmitter consists of a stable oscillator, usually a crystal oscillator, followed by a chain of amplifiers, one of which is modulated by one of the methods described in Chap. IX. The oscillator is practically never modulated because the accompanying frequency modulation would result in the production of high-order side bands. The modulation may take place in the final power amplifier, which is referred to as "high-level" modulation, or may take place in one of the lower power stages, which is termed "low-level" modulation. In the latter case it is necessary that all the amplifiers following the modulated stage be linear amplifiers in order to avoid distortion of the modulation. The use of low-level modulation has the advantage of requiring only a small amount of modulating power, but the disadvantage

rectifier-filter system supplying the final power tubes varies with the keying, thus setting up transients in the filter system which distort the transmitted impulses. For further information concerning such transients, see Reuben Lee, Radio Telegraph Keying Transients, *Proc. I.R.E.*, vol. 22, p. 213, February, 1934.

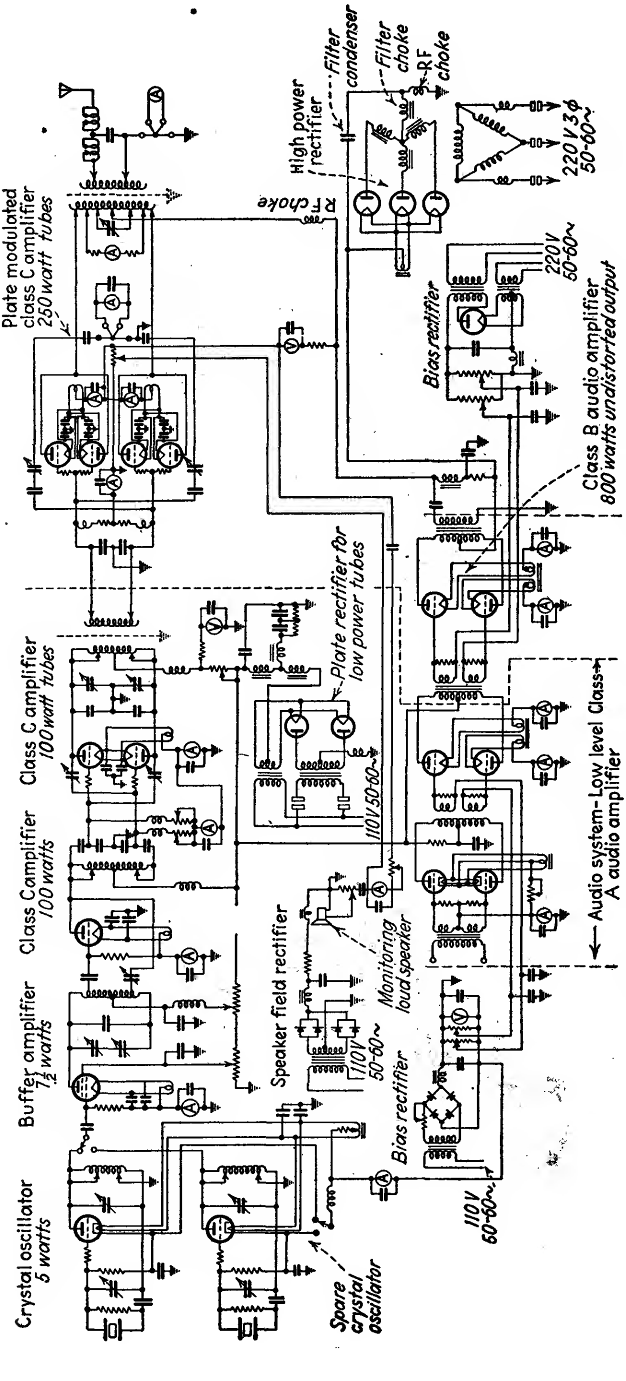
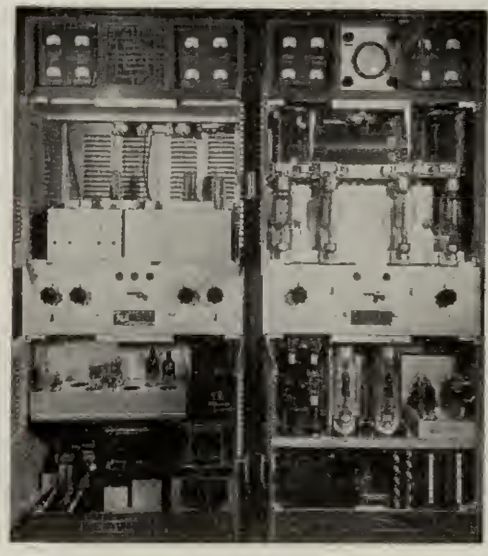


FIG. 288.—Simplified schematic circuit diagram of RCA 1-kw broadcast transmitter employing high-level modulation.

of requiring linear amplifiers, which are difficult to adjust and which also have low average efficiency when handling modulated waves. High-

1

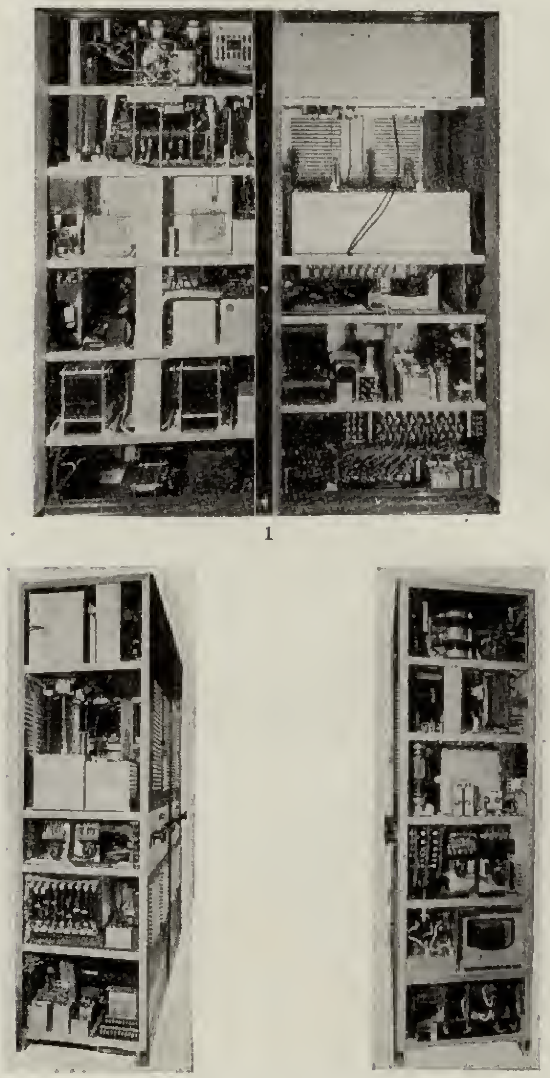


2

FIG. 289a.—Views of RCA 1-kw broadcast transmitter.

level modulation eliminates the necessity of linear amplifiers but requires correspondingly more audio-frequency modulating power. Both systems are capable of giving satisfactory performance, so that the choice is

determined by an economic balance between power costs, tube costs, etc., and this depends to a considerable extent upon the design details and the type of audio-frequency power amplifier employed.



1

2 3

FIG. 289*b*.—Views of RCA 1-kw broadcast transmitter.

Broadcast Transmitters.—The transmitters used in radio broadcasting represent the highest development in radio-telephone transmitters from the point of view of frequency stability, fidelity of modulation, etc.

Crystal-controlled oscillators are universally employed to meet the frequency-stability requirements, and, since the broadcast frequencies are in the band 550 to 1500 kc, it is possible to grind crystals to generate the desired carrier frequency and thereby to dispense with harmonic generators. The crystal oscillator is followed by one or more stages of Class C buffer amplifiers, after which comes the modulated amplifier. In the case of low-level modulation this is followed by one or more stages of linear amplification. The only methods of modulation employed to any extent in broadcast work are the grid-modulated and plate-modulated Class C amplifiers, both of which permit complete modulation with relatively little distortion.

Schematic diagrams of typical broadcast transmitters are shown in Figs. 288, 290, and 291. The transmitter of Fig. 288 is an RCA 1-kw model employing high-level modulation. The tube sequence in the radio-frequency chain is as follows:

5-watt crystal oscillator.
7½-watt screen-grid Class C amplifier.
100-watt neutralized triode Class C amplifier.
Two 100-watt triodes operated as neutralized push-pull Class C amplifiers.
Four 250-watt triodes connected in parallel push-pull, and plate modulated.

The audio-frequency system consists of a three-stage amplifier, the final stage of which is a Class B audio amplifier capable of delivering sufficient audio power (about 800 watts) to modulate completely the output of the final radio-frequency power amplifier. Photographs of this transmitter are shown in Fig. 289.

The schematic diagram of a 50-kw Western Electric transmitter using low-level modulation is shown in Fig. 290. The tube sequence in the radio-frequency system is as follows:

5-watt crystal oscillator.
5-watt Class C buffer amplifier.
50-watt Class C amplifier.
50-watt plate-modulated Class C amplifier.
Two 250-watt tubes as linear push-pull amplifiers.
Two 35-kw water-cooled tubes as linear push-pull amplifiers.
Six 35-kw water-cooled tubes as linear push-pull amplifiers.

The audio-frequency system consists of a two-stage amplifier, the final stage consisting of a 250-watt tube operated as a Class A amplifier. Complete modulation is obtained by operating the modulated radio-frequency amplifier at an appreciably lower plate voltage than the modulator tube.

The schematic circuit diagram of a 5-kw Western Electric transmitter employing low-level grid modulation is shown in Fig. 291. The sequence of tubes in the radio-frequency chain is as follows:

5-watt crystal oscillator.
5-watt buffer amplifier.
Two 5-watt tubes connected in parallel as Class C amplifier.
Two 250-watt tubes connected in push-pull and grid modulated.
Two 10-kw water-cooled tubes as push-pull linear amplifier.

The audio-frequency system consists of a two-stage amplifier, the second stage being a single triode capable of developing an undistorted power output of 10 watts when operating as a Class A amplifier. When the linear amplifier is omitted and the output of the modulator is delivered directly to the antenna, the result is a 100-watt transmitter suitable for low-power broadcast stations.

In comparing these three transmitters certain essential differences will be noted. The arrangement employing high-level modulation with a Class B audio amplifier has a greater over-all power efficiency and a smaller installed tube capacity than the other systems in proportion to carrier power and so tends to be preferred in high-power stations where the power and tube cost is high.[1] Low-level modulation, while less efficient than high-level modulation employing Class B modulators, has the advantage of requiring less audio-frequency amplification and of avoiding the possibility of audio distortion which develops in Class B audio amplifiers not carefully supervised. Grid modulation has the advantage of requiring relatively little audio power, but the disadvantage of making the modulator efficiency quite low.

The performance of a commercial broadcast transmitter leaves very little to be desired. Thus the manufacturers of the transmitter of Fig. 288 specify the following performance:[2]

Carrier stability ± 10 cycles.
Frequency response flat within 1 db from 30 to 10,000 cycles.
Audio harmonics (r.m.s.) less than 2.5 per cent at 400 cycles for 95 per cent modulation.

Short-wave Radio Telephone Transmitters.—Transmitters used for broadcasting at high frequencies are similar to the broadcast transmitters described above except for the fact that harmonic generators must be employed between the crystal oscillator and the modulator to obtain

[1] A description of a 500-kw broadcast transmitter employing high-level Class B audio modulation is described by J. A. Chambers, L. F. Jones, G. W. Fyler, R. H. Williamson, E. A. Leach, and J. A. Hutcheson, The WLW 500-kw Broadcast Transmitter, *Proc. I.R.E.*, vol. 22, p. 1151, October, 1934. This transmitter is notable in that the Class B audio amplifier required to modulate the 500-kw carrier completely consists of eight 100-kw water-cooled tubes developing an undistorted power output of approximately 800 kw.

[2] An extensive survey of the problem of high-quality broadcasting is given by Stuart Ballantine, High Quality Radio Broadcast Transmission and Reception, *Proc. I.R.E.*, vol. 22, p. 564, May, 1934.

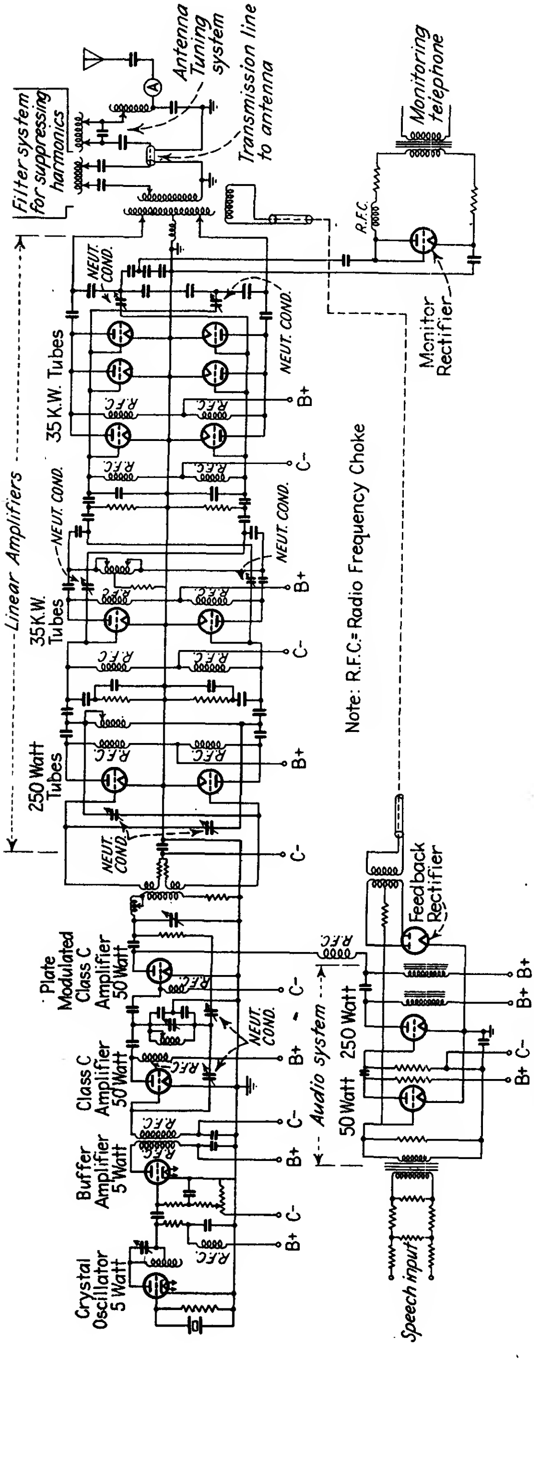


Fig. 290.—Simplified schematic circuit diagram of Western Electric 50-kw broadcast transmitter employing low-level modulation.

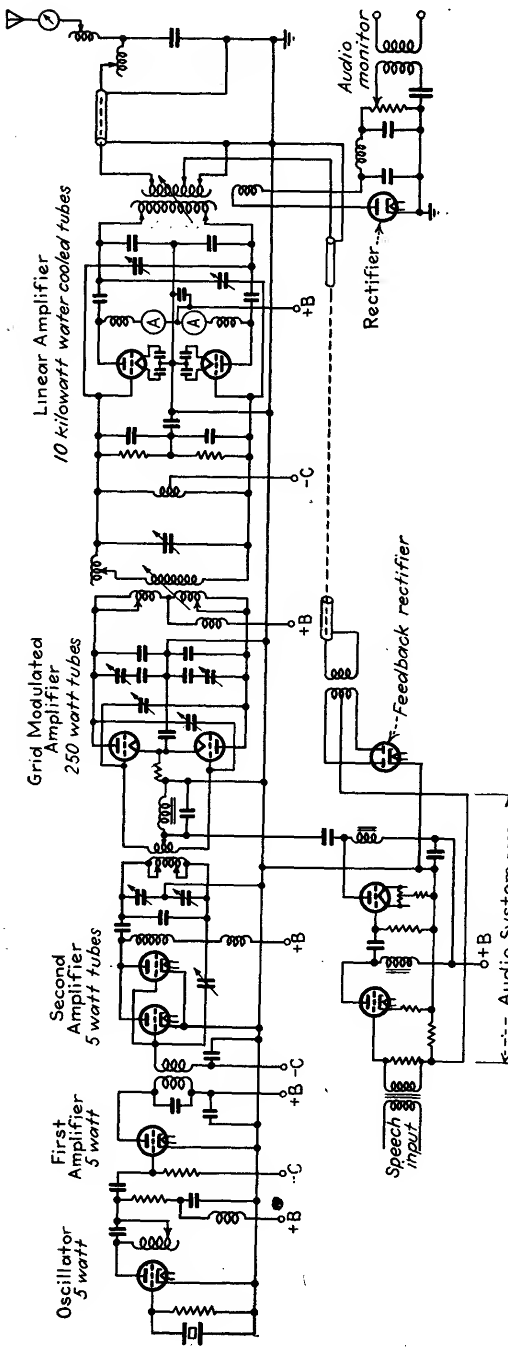


FIG. 291.—Simplified schematic circuit diagram of Western Electric 5-kw broadcast transmitter employing low-level grid modulation.

the high frequencies. Short-wave transmitters used for airplane, police, amateur, and other similar purposes are similar in a general way to short-wave broadcast transmitters but need not ordinarily meet the same high requirements with regard to stability of carrier frequency, audio-frequency range, and distortion. This permits the use of other modulation systems, such as the suppressor-grid modulated or control-

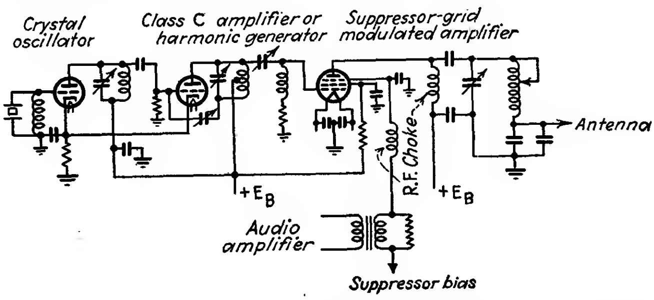


FIG. 292.—Typical amateur telephone transmitter employing suppressor-grid modulation.

grid modulated Class C amplifier in which the grid is driven positive. An example of a typical amateur transmitter employing suppressor-grid modulation is shown in Fig. 292.

Transmitters designed for use in airplanes must be light in weight, compact, and have the highest possible over-all power efficiency. This requires care in the design, and in some cases resort to special circuit arrangements to utilize effectively the full possibilities of the equipment.

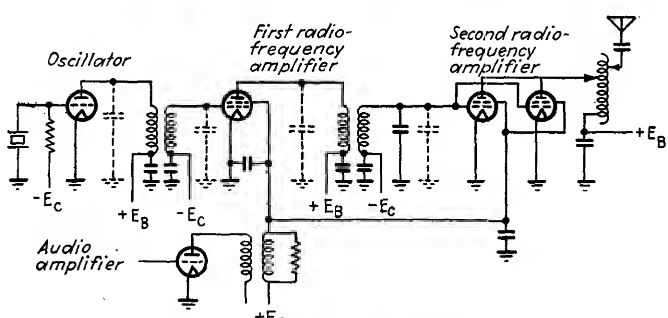


FIG. 293.—Schematic circuit diagram of Western Electric 50-watt transmitter intended for airplane use.

An example of such an arrangement is the Western Electric transmitter of Fig. 293. This consists of a crystal oscillator with the crystal ground to the frequency to be transmitted, followed by two stages of screen-grid radio-frequency amplification. Modulation is accomplished by simultaneously modulating the screen voltage of the tubes in both stages of radio amplification. The adjustment is such that the carrier wave is

partially modulated in the first radio-frequency amplifier tube, and this partially modulated wave is then completely modulated in the second

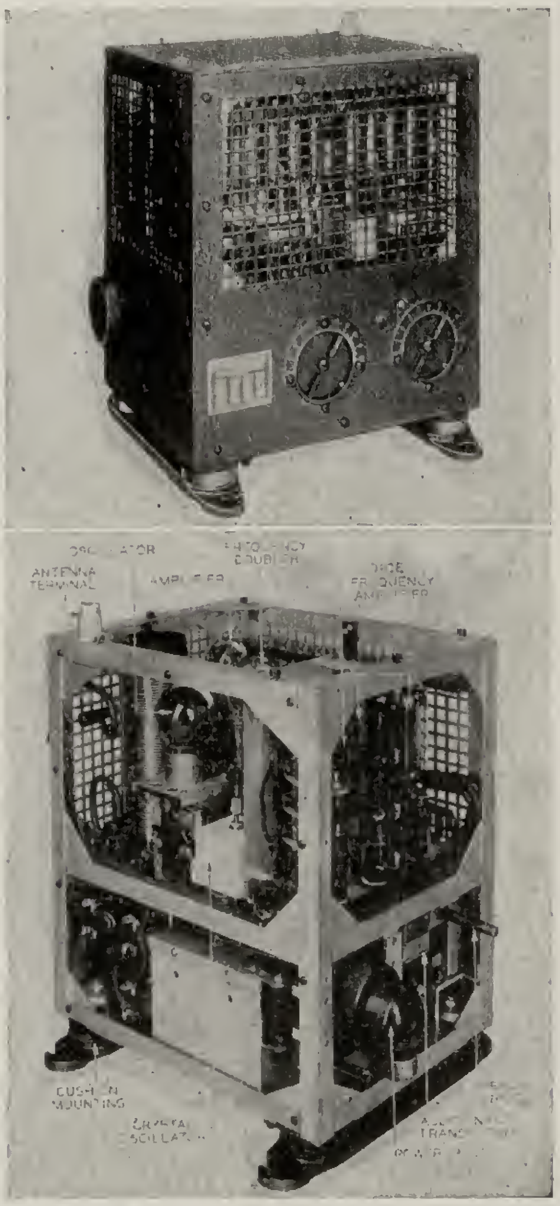


FIG. 294.—Photographs of typical airplane short-wave telephone transmitter.

radio-frequency amplifier. With this particular transmitter it is possible to obtain complete modulation of a 50-watt carrier with only 1 watt of

audio power and with a maximum distortion of less than 10 per cent. Photographs illustrating the kind of construction commonly employed with airplane equipment are given in Fig. 294.

Long-wave Radio-telephone Transmitters.—Radio-telephone transmitters operating below 500 kc are usually of the master-oscillator power-amplifier type with the modulation introduced into one of the Class C amplifiers. An example of such a transmitter is shown in the schematic circuit diagram of Fig. 295, which is a United States Coast Guard 500-watt transmitter intended for use in the frequency range 125 to 500 kc, and including provision for either telegraph or telephone communication. The transmitter consists of a 50-watt oscillator, an intermediate radio-frequency amplifier made of three 50-watt tubes in parallel, and a 1-kw

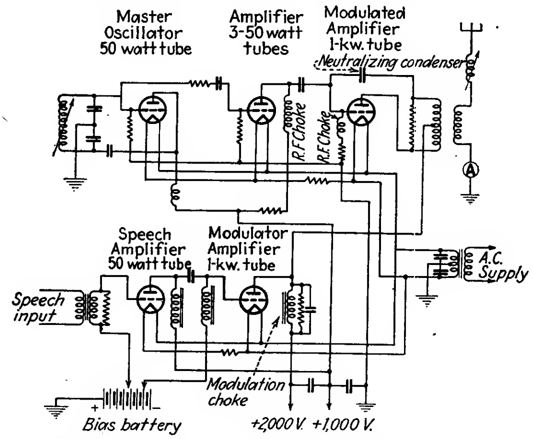


FIG. 295.—Circuit diagram of radio-telephone transmitter in which the frequency is controlled by a master oscillator. This transmitter is intended for use in the frequency range 125 to 500 kc, and is not arranged to give 100 per cent modulation.

output amplifier. Telephone communication is obtained by modulating the plate circuit of the output amplifier by a 1-kw modulator tube, the grid of which is excited by a 50-watt audio-frequency amplifier tube.

105. Miscellaneous Features of Radio Transmitters. *Control and Monitoring Systems.*—Large radio transmitters are usually made so that the transmitter may be placed in operation by pressing a single push button. The circuit thus closed causes the various contacts to be made in the proper sequence and with the requisite time delay, and interlocks are provided to prevent damage to the transmitter if some part fails to function. Transmitters are also provided with a certain amount of relay protection which will take care of overvoltages, undervoltages, failure of the circulating water in water-cooled tubes, short circuits, etc.

Satisfactory performance of radio transmitters is insured by systematic monitoring of the carrier frequency and the quality of the signals.

The frequency of broadcast transmitters is continuously checked against a secondary standard of frequency consisting of a carefully designed and operated crystal oscillator. The signals of other types of transmitters are normally checked periodically for frequency against a primary or secondary standard located within range of the transmitted signals.

In code transmitters the character of the transmitted signals obtained with hand keying is usually observed at the transmitting station by listening on an insensitive well-shielded radio receiver. In high-speed transmission, tape or oscillograph recording is used to detect faulty keying.

The usual provisions for monitoring the quality of signals transmitted by a radio-telephone transmitter comprise a loud-speaker operated by rectifying a portion of the modulated output with a linear detector, and some method, commonly a cathode-ray device, for observing the linearity and degree of modulation. In this way it is possible to maintain a close check upon the performance of the transmitter and to detect improper operating conditions such as overmodulation and other forms of distortion.

Hum in Radio-telephone Transmitters.—In broadcast transmitters trouble is commonly experienced from the alternating filament current modulating power-line hum upon the carrier in sufficient amount to cause annoyance during intervals when the carrier is otherwise unmodulated. Remedies include the use of direct-current generators for supplying filament power, the introduction of a neutralizing hum in the audio-frequency system, and the use of negative feedback as discussed below.

Application of Negative Feedback to Radio-telephone Transmitters.—The amplitude distortion and hum in the modulation envelope can be greatly reduced by making use of a modified form of negative feedback. One method of doing this is to apply a small portion of the radio-frequency output to a linear rectifier, thereby developing an audio-frequency voltage that reproduces the modulation envelope. This audio frequency is then fed back into the audio-amplifier system in the same way as a portion of the output is superimposed upon the input in the negative feedback amplifier discussed in Sec. 52. An alternative arrangement applicable to linear amplifiers but not to modulated amplifiers consists in simply coupling the input and output tuned circuits in such polarity as to give negative feedback. When the feedback process is properly carried out, the result in either case is essentially the same as in the ordinary feedback amplifier, there being a marked reduction in both alternating-current hum and in the distortion of the *modulation envelope*.

In applying negative feedback to radio-frequency systems it is necessary to take the same precautions as in ordinary negative feedback amplifiers, keeping in mind that a stage of tuned radio-frequency amplifi-

cation produces phase shifts in the modulation envelope which are of the same order of magnitude as the phase shifts produced in a single stage of audio-frequency amplification.

The transmitters shown in Figs. 290 and 291 make use of negative feedback. Here a portion of the output is rectified by a full-wave rectifier tube and is introduced into the grid circuit of the next-to-final audio stage in such a manner that the phase is correct for negative feedback.

Suppression of Harmonics.—Since Class C and linear amplifiers generate harmonics, it is necessary that care be taken to avoid the radiation of energy on frequencies other than the desired carrier frequency. The tank circuit of the final power amplifier helps greatly in suppressing the undesired harmonics, particularly if the effective Q of the circuit is high. However, in the case of high-power transmitters additional selectivity is required. This is commonly supplied by placing between tank circuit and antenna a coupling system involving inductances and capacities designed to discriminate very strongly against the harmonics of the carrier frequency, while transmitting the desired carrier and side bands with relatively little discrimination. The effectiveness of suppression obtained increases with the reactive volt-amperes circulating in the network, depends somewhat upon the extent the total inductance and capacity is subdivided, and is increased by the use of capacitive coupling.[1]

It is to be noted that the use of a push-pull connection will eliminate second harmonics provided the push-pull stage is balanced and the coupling between the tank circuit and the antenna is so arranged that stray capacities do not transfer second-harmonic energy.

Simultaneous Two-way Telephone Conversation on the Same Frequency. Radio-telephone channels that are links in wire telephone systems sometimes make use of the same frequency for transmission in both directions. This presents special problems because the radio receiver that picks up the signals from the distant transmitting station is necessarily in close proximity to the local transmitting station which is operated on the same frequency. Trouble is avoided by making use of relays which switch the land lines to the transmitter or to the receiver, depending on whether the speech is coming into, or going out of, the radio system. The normal relay position connects the receiving equipment to the land lines and disconnects the transmitters. If, however, voice currents are received at the radio station from the land line, a relay disconnects the

[1] For a detailed discussion of methods for avoiding harmonic radiation and particularly for information on the design of coupling networks, see J. W. Labus and Hans Roder, The Suppression of Radio-frequency Harmonics in Transmitters, *Proc. I.R.E.*, vol. 19, p. 949, June, 1931; Yuziro Kusunose, Elimination of Harmonics in Vacuum-tube Transmitters, *Proc. I.R.E.*, vol. 20, p. 340, February, 1932.

receiving equipment from the land lines and places the transmitter on the circuit instead. In order that no speech will be lost during the brief interval when the relays are changing the connections, a time delay provided by an artificial line or a long speaking tube is placed between the point in the circuit where the voice energy operates the relay and the point where the relay contacts change the connections. A schematic diagram of the circuits involved is shown in Fig. 296.[1]

Production of Single Side-band Signals.—Radio-telephone signals consisting only of a single side band with the carrier and companion side band suppressed have certain advantages. Such signals are produced by initially modulating the audio-frequency signals upon a relatively low carrier frequency using the balanced modulator system of Fig. 231 to suppress the carrier. The undesired side band is then eliminated by the use of filters, which can be made sufficiently selective to separate one side band from the other provided the carrier frequency is not too great. The remaining side band is then modulated upon a second carrier wave of considerably higher frequency than the first, and one of the resulting side bands is again selected by the use of suitable filters. This process can be repeated as many times as is necessary to bring the frequency up to the desired value, after which linear amplification is employed to produce the required power. The only single side-band radio system in commercial operation is the long-wave transatlantic telephone which has the schematic transmitter layout illustrated in Fig. 297.[2]

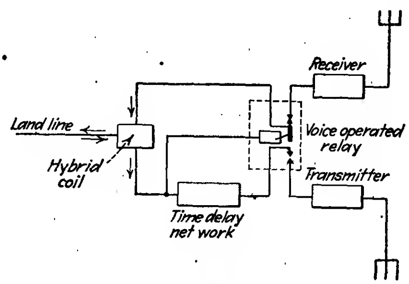


FIG. 296.—Circuits involving voice-operated relay and time-delay network for making possible two-way radio-telephone communication using the same frequency for transmission in both directions.

Synchronous Operation of Broadcasting Transmitters.—In a number of instances the carrier frequencies of two or more broadcast stations have been synchronized in order to reduce interference. The usual method of doing this involves the use of a relatively low control frequency transmitted over telephone wires. For example, in one installation a 4-kc

[1] Arrangements of this type are used in most radio links which attempt to provide two-way talking service, even when different transmission frequencies are used in the two directions. For example, see G. C. Crawford, Echo Elimination in Trans-atlantic Service, *Bell Lab. Record,* November, 1927.

[2] A corresponding single side-band short-wave radio-telephone system has been devised and used experimentally for transatlantic operation with considerable success. See F. A. Polkinghorn and N. F. Schlaack, A Single Side Band Short-wave System for Transatlantic Telephony, *Proc. I.R.E.*, vol. 23, p. 701, July, 1935.

signal is transmitted to each broadcast station and is used as a reference against which the frequency actually transmitted is continuously compared by automatic means. If the transmitted frequency deviates from synchronism with a harmonic of the reference frequency, means are provided by which the transmitted frequency is automatically brought back to the desired value. A number of variations of this general procedure are possible, such as transmitting the control frequency by radio, continuously comparing the two transmitted frequencies against each other and then either automatically or manually adjusting one of the transmitters if the frequency difference becomes excessive, etc.[1]

Adjacent Channel Interference.[2]—Radio-telephone transmitters sometimes produce interference on frequencies that are outside of the normal side-band range and yet are relatively close to the carrier frequency. Such adjacent channel interference is caused by high-order side bands resulting either from distortion of the modulation envelope or from

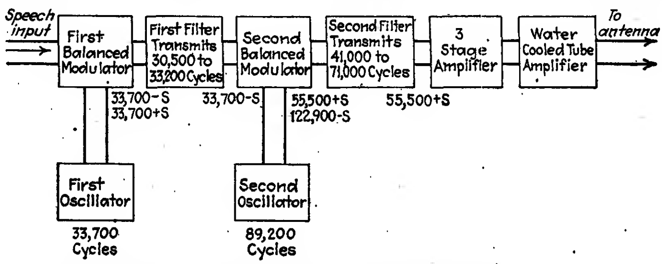


FIG. 297.—Schematic layout of long-wave transatlantic telephone, showing how the single side band is generated and amplified to a high-power level.

frequency or phase modulation of the transmitter. Thus, if the highest modulation frequency supplied to a transmitter is 5000 cycles, distortion of the modulation envelope produces second-order side bands that extend to 10,000 cycles on either side of the carrier, fifth-order side bands extending 25,000 cycles, etc. These extended side bands will commonly interfere with the local reception of signals from distant transmitters long before the distortion involved is of sufficient magnitude otherwise to impair the transmitter performance. Overmodulation caused by occasional audio-frequency peaks is the commonest cause of adjacent channel interference, since, when the peak value of the modulating voltage becomes greater than that required for 100 per cent modulation,

[1] For further information on synchronized broadcast transmitters, see L. McC. Young, Present Practice in the Synchronous Operation of Broadcast Stations as Exemplified by WBBM and KFAD, *Proc. I.R.E.*, vol. 24, p. 433, March, 1936; G. D. Gillett, Some Developments in Common Frequency Broadcasting, *Proc. I.R.E.*, vol. 19, p. 1347, August, 1931.

[2] For further information on this subject, see Ira J. Kaar, Some Notes on Adiacent Channel Interference, *Proc. I.R.E.*, vol. 22, p. 295, March, 1934.

the trough of the modulation envelope will be zero for a portion of the time, giving rise to a distortion that ordinarily contains high-order harmonic components.

Phase and frequency modulation produce high-order side bands as explained in Sec. 81, and hence can give rise to adjacent channel interference. Frequency modulation occurs when the oscillator is modulated, or when the buffer amplifier located between the oscillator and the modulated stage allows the modulation to produce a reaction back upon the oscillator. Phase modulation occurs when the tank circuit of the modulated amplifier or one of the succeeding linear amplifiers is not tuned exactly to resonance, or when there is some coupling between circuits carrying modulated currents and circuits on the oscillator side of the modulated tube which carry unmodulated currents.

Transmitters Employing Frequency and Phase Modulation.—Frequency and phase modulation are similar to each other insofar as the end result is concerned, but differ in the method by which the modulation is obtained and in the effect of the modulating frequency upon the modulation index.

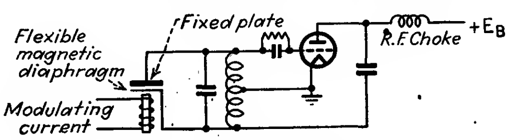


FIG. 298.—Simple arrangement for producing a frequency-modulated signal.

The term *frequency modulation* is hence commonly used to designate phase modulation as well as pure frequency modulation. No commercial communication system using frequency modulation has yet been placed in operation, but extensive experimental work upon such systems is being carried out, and these systems show promise for communication at very high frequencies.

Pure frequency modulation is obtained by varying the generated frequency in accordance with the amplitude of the modulated voltage. This can be accomplished in various ways, a typical arrangement being illustrated in Fig. 298. Here the modulating energy actuates an electromagnet that displaces a diaphragm representing one plate of a condenser that is part of the tuning capacity of the oscillator. A simple method of obtaining phase modulation is illustrated in Fig. 299*a*. Here a portion of the energy from a master oscillator is amplitude-modulated in any convenient manner, and after being shifted in phase by 90° is then combined with a larger amount of unmodulated energy from the same master oscillator. The resulting wave has a phase and amplitude that vary with the modulation as illustrated by the vector diagrams of Fig. 299*b*, and after being passed through a limiting amplifier to remove the residual amplitude modulation represents a phase-modulated wave.

The modulation index obtained with frequency or phase modulation is increased by the use of harmonic generators. Thus, if the original phase modulation represented a 20° variation in phase, three successive frequency doublers would increase the phase variation to $2 \times 2 \times 2 \times 20 = 160°$.

Secrecy Systems.—In radio extensions of telephone systems it is desirable to preserve privacy. One possible method of doing this consists in inverting all the speech frequencies by transmitting the high speech frequencies as low modulation frequencies and the low audible frequencies as high modulation frequencies. This can be accomplished by heterodyne action. Thus, if the frequency band to be transmitted is 400 to 3000 cycles, heterodyning the speech band with 3400 cycles and then separating out the difference frequencies by means of a suitable filter will transform an original 3000-cycle sound to 400 cycles, an original 400 cycles to 3000 cycles, and so on. At the receiving end the speech can be restored by a second heterodyne operation. The system can be made

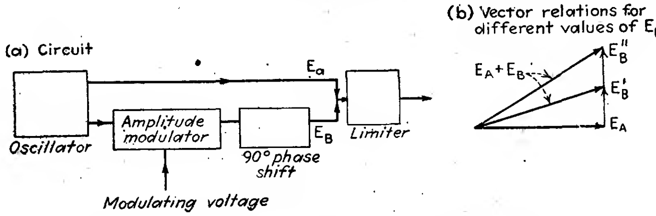


FIG. 299.—Method of generating a phase-modulated wave.

still more difficult to decipher by allowing the heterodyning frequency to wobble slowly about 3400 cycles and then to wobble the heterodyning oscillator at the receiver correspondingly. Other secrecy systems involve such operations as dividing the audible range into bands such as 400 to 1000 cycles, 1000 to 1600 cycles, etc., and interchanging these bands by making use of the heterodyne principle. If, in addition, means are provided for automatically changing the method of rearrangement every few moments according to a prearranged system, complete privacy is practically assured.

Interrupted Continuous Waves.—In code transmitters it is sometimes desired to interrupt the transmitted dots and dashes at a rate approximately 1000 cycles per second, giving what is termed I.C.W. (interrupted continuous waves). This is commonly accomplished by the use of a self-rectifying circuit using 500-cycle power or by the use of some form of motor-driven interrupter or chopper. Interrupted continuous waves can be received upon ordinary radio receivers without heterodyne action; they also have the advantage in short-wave work that a certain amount of frequency diversity is obtained as a result of the side bands produced by the interruptions.

Size of Successive Tubes in the Radio-frequency Chain.—The size of successive tubes in the chain of radio-frequency amplifiers is determined by the fact that the final amplifier tube must deliver the required output, while each preceding tube must deliver sufficient output to excite the tube that follows it. Inasmuch as the driving power required by a particular tube depends greatly upon the exact operating conditions, the size of tube required for excitation will depend greatly upon the design details. Furthermore, the exciting power required by different tubes of the same type will often vary considerably as a result of variations in secondary electron emission at the control grid. Consequently it is usually desirable to design conservatively and to provide considerably more driving power at each stage than is expected to be required under ordinary conditions.

Transmitters for Very High Frequencies.—At frequencies exceeding about 20 megacycles crystal-oscillator arrangements become complicated because of the large number of harmonic generators required. Under such circumstances ordinary oscillators, in which the frequency is made stable by employing a specially designed resonant circuit having an extremely high Q, are commonly employed. The most widely used arrangement of this type makes use of a resonant transmission line to control the frequency as discussed in Sec. 67.[1] When temperature compensation is incorporated, the frequency stability compares favorably with that of crystal-controlled transmitters.

For wave lengths down to about ½ to 1 meter it is still possible to employ master-oscillator power-amplifier arrangements, the preferable arrangement being to develop the final output power from a harmonic generator.[2] At shorter wave lengths it is necessary to employ electron oscillators either of the Barkhausen or split-anode type. Means have been devised for modulating the output of such oscillators so that radio-telephone signals can be transmitted, although in general these modulation systems all introduce an undesirably large amount of frequency modulation.

106. High-efficiency Systems for Radio-telephone Transmitters.—None of the layouts that have been described for radio-telephone transmitters are ideal. The high-level plate-modulated systems require a

[1] Another form of high-frequency oscillator having a high Q-resonant circuit is described by F. A. Kolster, Generation and Utilization of Ultra-short Waves in Radio Communication, *Proc. I.R.E.*, vol. 22, p. 1335, December, 1934. This employs a specially designed one-turn inductance so arranged that the losses in the inductance and in the associated tuning capacity are extremely low. With temperature compensation such an arrangement compares favorably with the resonant-line oscillator, although it is not so satisfactory structurally.

[2] See N. E. Lindenblad, Development of Transmitters for Frequencies above 300 Megacycles, *Proc. I.R.E.*, vol. 23, p. 1013, September, 1935.

very large amount of undistorted audio-frequency power, which is difficult to obtain, and high-level grid modulation, while requiring relatively little audio power, is very inefficient. Low-level modulation on the other hand requires the use of linear amplifiers, which have low average efficiency when designed to handle a completely modulated wave. In order to overcome or minimize these disadvantages, a number of high-efficiency systems as described below have been devised.

Controlled-carrier Transmitters.[1]—This type of transmitter employs low-level modulation so arranged that the carrier is turned on and off as the amplitude of modulation varies. The transmitted wave in a typical case is therefore as illustrated in Fig. 300, with the carrier amplitude varying in accordance with the envelope of the modulating voltage. When such a wave is applied to a linear amplifier, the average loss is

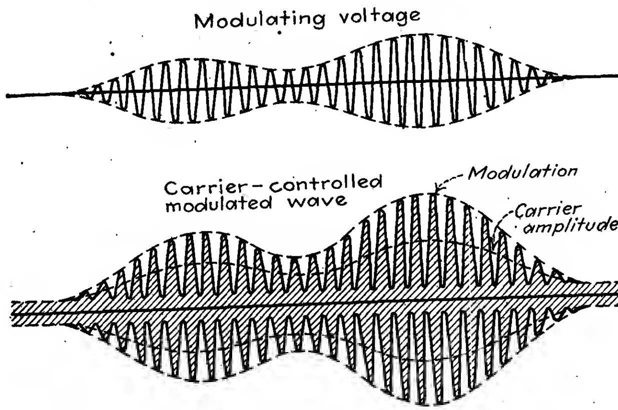


FIG. 300.—Wave transmitted from a station employing controlled carrier.

low because during the silent intervals little or no excitation is applied to the linear amplifier and the tube losses are low, while during the peaks of modulation the linear amplifier is developing full output but is then operating under conditions favorable for high efficiency. Furthermore, with ordinary sounds, the average amplitude is much less than the peaks, which are also relatively infrequent. The satisfactory reception of a carrier-controlled signal presupposes the use of automatic volume control of suitable characteristics at the receiving point in order that the output will be substantially independent of the carrier amplitude.

A typical circuit for producing a modulated wave having controlled carrier is illustrated in Fig. 301. Here low-level plate modulation using a Class B audio modulator is employed. The control of the carrier is

[1] For further details concerning controlled-carrier transmitters, the reader is referred to recent issues of *QST* and *Radio*. Condensed information on the same subject is to be found in "Radio Amateur's Handbook" and "The Radio Handbook."

obtained by using a separate power-supply system for the plate circuit of the modulated tube, and varying this supply voltage in accordance with the modulation envelope by using the d-c current in the plate circuit of the Class B tubes to control a saturable reactor in series with the primary circuit of the power-supply system. Since the d-c plate current of the Class B audio tubes is proportional to the amplitude of the modulating voltage, the saturation of the saturable reactor will vary likewise, causing the voltage across the primary of the power transformer to increase as the modulation comes on and to decrease during the quiet intervals.

The controlled-carrier transmitter has the advantage of enabling a given linear amplifier to develop nearly twice as much power output for a

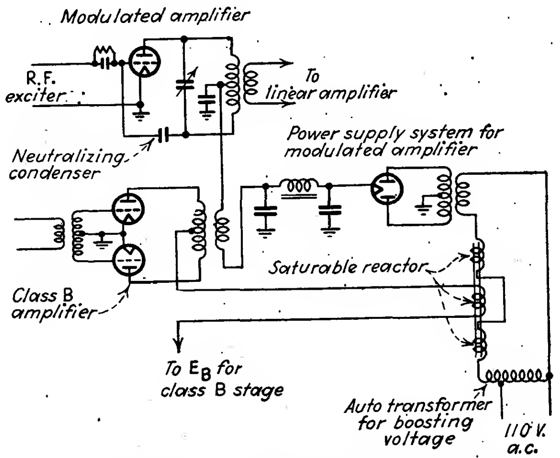


FIG. 301.—Circuit of controlled-carrier transmitter, showing saturable reactor method of control.

given average dissipation as when an ordinary modulation system is employed. The chief disadvantage is that, unless some form of time delay is employed, there is a tendency for distortion to occur when the amplitude of modulation is suddenly increased, since the amplitude of the carrier cannot be increased instantly. Furthermore it is necessary that the receiver have automatic volume control with suitable time constants. The result is that, while the controlled-carrier system of transmission is not suitable for broadcast use, it has found considerable favor in amateur work and in other transmitters where a limited amount of distortion can be tolerated and where economy is of importance.

High Efficiency Obtained by Dynamic Shift of the Operating Conditions.[1]—The reason for the low average efficiency of linear amplifiers is

[1] This method of obtaining high efficiency was developed independently but more or less simultaneously by the author and J. N. A. Hawkins. For further details see

that, when the carrier is unmodulated, the adjustments must be such that the minimum plate voltage of the linear amplifier exceeds half the direct-current supply voltage. This is necessary to accommodate the peak amplitudes occurring with complete modulation. The average efficiency can hence be improved by allowing the direct-current plate voltage to vary in accordance with the envelope of the modulating voltage in such a way that at the peak of the modulation the direct-current plate voltage of the linear amplifier is large, while during unmodulated intervals the direct-current plate voltage is reduced to 65 per cent or thereabouts of its peak. By varying the grid bias along with the plate voltage so that the operating

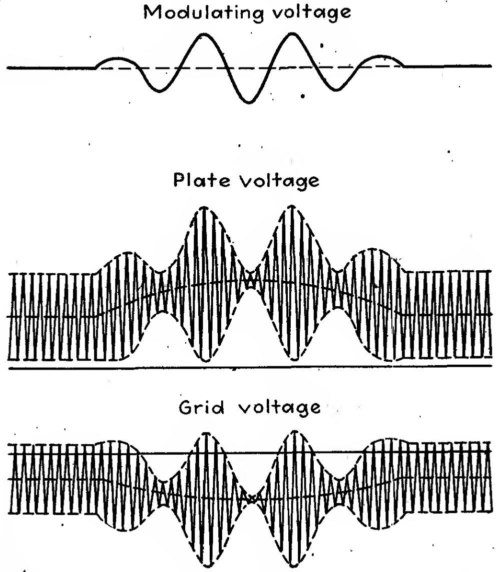


FIG. 302.—Oscillograms illustrating the action of a linear amplifier operated at high efficiency by dynamic shift of the operating point.

point is always maintained at the same cut-off or projected cut-off point, the linearity of the amplifier will not be interfered with. Oscillograms illustrating the operation are shown in Fig. 302.

A typical schematic circuit for carrying out the necessary operations is illustrated in Fig. 303. Here a saturable reactor is used to control the voltages applied to the grid and plate electrodes of the linear amplifier. The d-c current for this reactor is obtained by rectifying a portion of the audio-frequency signal, so that the saturation will therefore vary with the

J. N. A. Hawkins, A New High-efficiency Linear Amplifier, *Radio*, no. 209, p. 8, May, 1936; Expanding Linear Amplifier Notes, *Radio*, no. 210, p. 63, June, 1936; F. E. Terman and F. A. Everest, Dynamic Shift, Grid Bias Modulation, *Radio*, no. 211, p. 22, July, 1936.

envelope of the modulation envelope. Only a portion of the grid bias is obtained from the variable source, and the ratio between fixed and variable bias is so proportioned that, as the saturation of the reactor is varied, the ratio of plate-voltage change to grid-voltage change is equal to the amplification factor of the tube, thereby maintaining the desired operating condition irrespective of the saturation of the reactor. In such an arrangement the plate-supply voltage during the quiet intervals is just sufficient to accommodate the unmodulated carrier with a reasonable margin of safety. As the modulation comes on, the reactor saturates, increasing both grid and plate voltages and thereby enabling the linear amplifier to accommodate the increased amplitude of the exciting voltage.

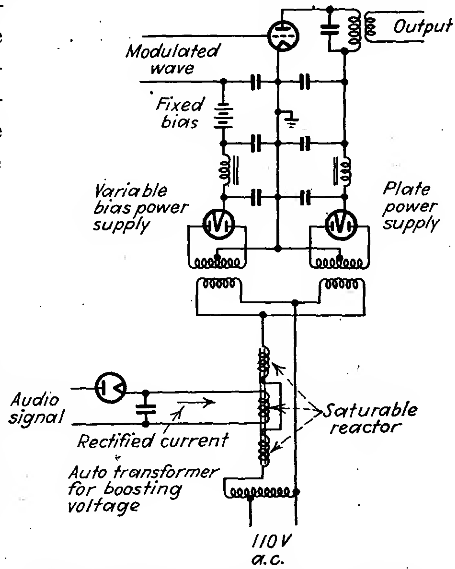


FIG. 303.—Schematic circuit for high-efficiency linear amplifier employing dynamic shift of the operating point.

These principles can also be applied to obtain a high-efficiency grid or suppressor-modulated amplifier. It is merely necessary to readjust the relative proportions of variable and fixed grid bias so that, when the exciting voltage is present, but with no modulating voltage, variation in the saturation of the reactor will not produce appreciable change in the carrier output. The modulation is then superimposed upon the control-grid or suppressor as the case may be.

As compared with controlled carrier, the dynamic-shift method of obtaining high efficiency has the advantage of higher efficiency and of providing a carrier of constant amplitude. The chief disadvantages are the accessory equipment required to produce the dynamic shift and the fact that, unless time delay is introduced in the audio-frequency circuits leading to the modulator, there will be a tendency toward distortion when the modulation suddenly increases, exactly as in the case of the controlled carrier.

High-efficiency Linear Amplifier.—It is possible to maintain the average efficiency of a linear amplifier at a value exceeding 60 per cent by an ingenious arrangement due to Doherty.[1] A schematic diagram of this amplifier is illustrated in Fig. 304*a*. Here the amplifier is divided into two

[1] For further information see W. H. Doherty, A New High Efficiency Power Amplifier for Modulated Waves, *Proc. I.R.E.*, vol. 24, p. 1163, September, 1936.

parts: A_1, which delivers its output to the load through an artificial line equivalent to a quarter-wave-length transmission line, and A_2 which

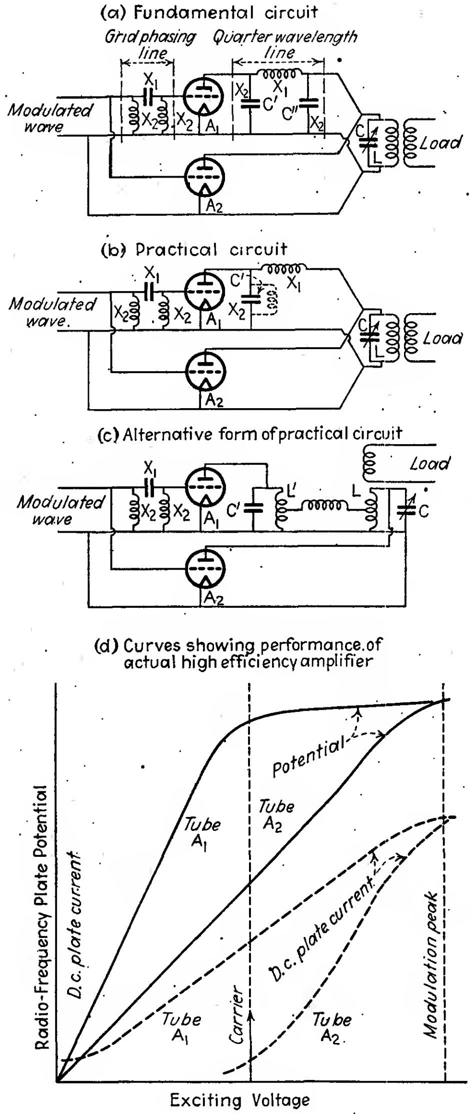


FIG. 304.—Schematic diagram of Doherty's high-efficiency linear amplifier.

delivers its output directly to the load. A phase shift of 90° is provided in the excitation of one of the amplifiers to compensate for the 90° phase shift in the artificial line and to allow the outputs of the two amplifiers to

add. Amplifier A_1 is adjusted to operate as an ordinary linear amplifier while A_2 is biased so that no plate current flows until the exciting voltage exceeds the value corresponding to the carrier amplitude.[1]

The circuits are so designed that, when the carrier voltage is applied to A_1 and there is no plate current flowing in A_2, the effective load impedance in the plate circuit of amplifier A_1 is such as to make the minimum plate voltage of this tube quite small, while at the same time the alternating-current voltage across the load impedance (*i.e.*, in the plate circuit of A_2) is exactly half the alternating-current voltage in the plate circuit of A_1. If the load impedance required in the plate circuit of A_1 is a resistance R_1, then the required relations are realized by making the load impedance actually used in the plate circuit of A_2 a resistance $R_1/4$ and by designing the quarter-wavelength line to have a characteristic impedance $R_1/2$.[2] With this arrangement the amplification is substantially linear, while the efficiency normally exceeds 60 per cent.

The operation can be explained as follows: For exciting voltages that do not exceed the carrier amplitude, amplifier A_2 is inoperative and A_1 functions as an ordinary linear amplifier developing an output voltage proportional to the exciting voltage. When the exciting voltage exceeds the carrier amplitude, however, A_2 also supplies energy to the load. This causes the equivalent load impedance to the A_2 end of the transmission line to be increased because the output from A_2 increases the voltage across the load in proportion to the current delivered to the load by the line. Now in a quarter-wave-length transmission line the impedance at the sending end of the line is equal to Z_0^2/Z_2, where Z_0 is the characteristic impedance of the line and Z_2 is the equivalent load impedance at the receiving end. Consequently, as amplifier A_2 delivers energy to the load, the impedance which the sending end of the line presents to the plate of A_1 is decreased, thereby enabling A_1 to deliver more output power with the same alternating-current voltage in the plate circuit, until at the peak of a completely modulated wave each tube delivers to the load twice as much energy as the carrier power.

The details involved are illustrated in Fig. 304*d* for an actual case. It will be noted that the alternating voltage in the plate circuit of tube A_1 increases almost linearly with excitation up to the carrier level and then flattens off abruptly at a value for which the crest alternating voltage is

[1] The bias voltage required to accomplish this is very nearly equal to half the bias required for cut-off with the plate-supply voltage used, plus an added bias voltage equal to the crest amplitude of the exciting carrier voltage. This is arrived at by noting that the alternating voltage in the plate circuit of amplifier A_2 when there is no modulation is just slightly less than half the direct-current plate-supply voltage.

[2] The series reactance X_1 and shunt reactance X_2 required in the artificial lines of Figs. 304 and 305 to make the length exactly a quarter wave length and to have a characteristic impedance Z_0 are $X_1 = X_2 = Z_0$.

only slightly less than the battery potential. The d-c plate current and hence the input power do not flatten off, however, but continue to increase almost linearly with the exciting voltage up to an amplitude corresponding to the peak of the modulation cycle. The alternating voltage in the plate circuit of amplifier A_2 (*i.e.*, the output voltage) is substantially linear up to the peak of the modulation cycle, while the d-c current of the second tube does not begin to flow until the exciting voltage approaches the carrier value, but it then increases rapidly and becomes equal to the d-c current of amplifier A_1 at the peak of the modulation cycle.

Tube A_1 operates with low plate losses and good efficiency because, when the exciting voltage equals or exceeds the carrier value, the minimum plate voltage is small, while, with exciting voltages less than the carrier value where the efficiency is low, the power input and hence the plate losses drop off rapidly. Tube A_2 likewise operates with low plate losses and good efficiency because no current flows in the plate circuit until the minimum plate potential is approximately half the supply voltage, and then there is not much input to the tube until the crest of the modulation cycle is approached, when the tube operates as a high-efficiency Class C amplifier. As a consequence the over-all efficiency in practice is of the order of 60 to 65 per cent, and is substantially independent of the degree of modulation.

The average power developed in the high-efficiency linear amplifier is divided unevenly between the tubes. With no modulation A_1 develops the full carrier output and A_2 has no output, while with full modulation the powers developed by A_1 and A_2 are, respectively, 0.93 and 0.57 of the carrier power. Hence the output of A_1 is roughly constant at the carrier value, while the output of A_2 is approximately equal to the side-band power. The division of losses is still more uneven, since A_2 operates on the average at higher efficiency than A_1, and also is called upon to develop power only in proportion to the modulation. The result is that, although A_2 must develop as much power to the peak of the modulation cycle as A_1, the plate dissipation that must be provided in A_2 is much less.

In practical high-efficiency amplifiers it is permissible to combine the receiver-end capacity C'' of the line with the tuned load circuit, while a low-impedance path to harmonics in the plate circuit of A_1 can be obtained by substituting for the capacity at the A_1 end of the line a sharply resonant circuit detuned just enough to offer the required capacitive reactance, as in Fig. 304*b*. An alternative arrangement which permits the use of a transmission line with low characteristic impedance and hence low X_1 is shown in Fig. 304*c*. In arranging for the excitation of amplifiers A_1 and A_2 it is commonly found that, with an adjustment that causes the carrier input to develop the proper output in A_1, the peak of the modulation

cycle will cause the grid of A_1 to be driven so far positive that the grid current and hence the driving power will be excessive. This trouble can be avoided by obtaining part of the bias for A_1 from a grid-leak arrangement so that the bias will increase at the peak of the modulation cycle, or it can be avoided by using a phasing arrangement consisting of a quarter-wave-length line arranged as shown in Fig. 304*b*. In this latter arrangement, increased grid current reduces the impedance at the grid end of the line and thereby lowers the voltage at the peaks, and incidentally reduces the driving power.

In the practical adjustment of a high-efficiency linear amplifier the 90° phase shift required between the grids of the two tubes can be realized with the aid of a cathode-ray oscillograph. In tuning up the plate circuit of the arrangement shown in Fig. 304*a* and *b*, the usual procedure is to make the line inductance X_1 the calculated value and then to adjust the tuned load circuit LC until a cathode-ray oscillograph shows a 90° phase difference at the two ends of the line. The condenser C' at the sending end of the line is then adjusted for minimum plate current in A_1 when the exciting voltage is such as to make A_2 inoperative. The load is then coupled into the tank circuit LC and adjusted until the voltage across the load is half the voltage across the sending end of the quarter-wave-length line when the excitation is such as to make A_2 inoperative. With the circuit of Fig. 304*c* the adjustment procedure is the same except for the fact that the point at which the transmission line is tapped to the tuned circuits affects the load coupling required and also the tuning adjustments of condensers C and C'.

This type of amplifier, while very new, shows every prospect of becoming important in cases involving high power. The over-all efficiency obtainable is higher than with any other system, even including high-level Class B audio modulation, while at the same time the advantages of low-level modulation are retained. The chief disadvantage is that, since the artificial line is not exactly a quarter wave length long to the side-band frequencies, the system does not function exactly as outlined, particularly when the modulation frequency is high, with consequent distortion at high modulation frequencies.

High-efficiency Grid- and Suppressor-modulated Amplifiers.—The same principles outlined above can be applied to grid- or suppressor-modulated amplifiers to obtain high-efficiency high-level modulation with low audio power. A schematic circuit diagram is illustrated in Fig. 305, and is essentially the same as for the linear amplifier of Fig. 304 except that the tubes, instead of being excited with a modulated wave, are excited by an unmodulated carrier voltage, with a modulating voltage applied to either control or suppressor grids. Amplifier A_1 is operated as an ordinary grid- or suppressor-modulated tube with the circuit adjustments such that with

no modulation and the second amplifier inoperative the minimum plate voltage will be small, while the alternating-current voltage in the plate circuit of A_2 is half the voltage in the plate circuit of A_1. The second amplifier is so biased that it delivers output only on the positive half cycles of the modulation. Under these circumstances the tubes divide the load exactly as in the case of the high-efficiency linear amplifier, and the average efficiency is approximately the same, *i.e.*, at least 60 per cent.

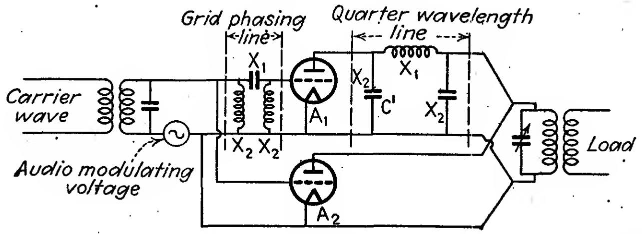


FIG. 305.—Schematic diagram of high-efficiency grid-modulated amplifier

Outphasing System of Modulation.—This is an ingenious arrangement which takes advantage of the fact that, when the side-band frequencies are shifted 90° from the phase position existing in an amplitude-modulated wave, the envelope of the resulting wave is of substantially constant amplitude irrespective of the amount of side band present.[1] Such a wave can be amplified without distortion and with high efficiency by

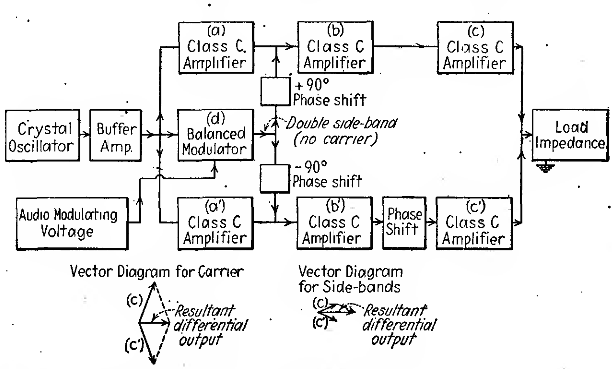


FIG. 306.—Schematic layout of transmitter employing outphasing modulation.

a Class C amplifier, after which it can be converted back into an amplitude-modulated wave by a suitable phase shift of the carrier with respect to the side-band components.

[1] The wave that results from this shift in phase of the side-band components is practically identical with a phase-modulated wave, the only difference being that there are no second- and higher order side bands present. The result can therefore be termed a quasi phase-modulated wave.

A practical arrangement for carrying out the necessary operations is illustrated schematically in Fig. 306. The balanced modulator shown can be of the type illustrated in Fig. 231 and produces side-band frequencies substantially free of carrier. A portion of the output of this modulator is shifted $+90°$ in phase and combined with the output of the Class C amplifier marked a, while another portion is shifted in phase $-90°$ and combined with the output of the Class C amplifier marked a'. These resultant outputs are then amplified to the desired power level, and provision is made for shifting slightly the relative phases of the two amplifiers. The outputs of the two final Class C amplifiers marked c and c' in Fig. 306 are then connected to the load so that the input to the load is the vector difference of the outputs of the amplifiers. With this arrangement the carrier supplied to the load is the difference of two equal vectors which are almost but not quite in phase opposition, and so is almost 90° out of phase with the carriers in the individual amplifiers as shown by the vector diagram of Fig. 306. On the other hand the side-band frequencies combine substantially in phase in the output because they were initially in opposite phase in the two amplifier branches. As a result, the original modulated wave is regained in the output.

The outphasing system of modulation combines the advantages of low-level modulation with the high efficiency of Class C amplification, and with proper adjustment it will give very low distortion. The method has been used commercially in several of the high-power French broadcasting transmitters with entirely satisfactory results.[1]

Problems

1. Design a code transmitter delivering 300 watts output at a frequency of 7500 kc and having crystal control with the crystal operating at half the transmitted frequency. The design includes a complete circuit layout with circuit constants specified as far as possible, choice of tubes, and rough determination of tube-operating conditions (such as preliminary estimates of grid bias, driving power, d-c plate current, plate efficiency, etc., for each stage). In this problem make full use of tube data given in tube manuals, and provide approximately 100 per cent more driving power than estimates indicate is necessary.

2. Explain in detail the operation of the keying system used in Fig. 287.

3. Explain why the transmitter of Fig. 282 does not produce serious key clicks.

4. *a.* Make sketches of the systems used in the transmitters of Figs. 288, 290, and 291 to couple the modulator tube to the modulated amplifier.

b. In each of these arrangements explain the factors that determine the extent to which the high and low audio frequencies fall off as a result of imperfect coupling.

5. In the transmitter of Fig. 288, draw to a large scale the circuit of the output stage (the part to the right of the dotted line), label each circuit element, and explain its purpose.

[1] For further information see H. Chireix, High Power Outphasing Modulation, *Proc. I.R.E.*, vol. 23, p. 1370, November, 1935.

6. In the transmitter of Fig. 288 determine values for the filter inductance and condenser such that the ripple voltage modulated upon the transmitter output will not produce a degree of modulation exceeding 0.001.

7. In the transmitter of Fig. 290, explain (*a*) the purpose of the resistance shown in series with the neutralizing condenser of the 50-watt modulated tube, (*b*) the reason for the resistance between ground and the center of the tank-circuit inductance of the first linear amplifier employing 35-kw tubes.

8. Design a crystal-controlled transmitter for operation at 1600 kc and capable of developing 50 watts of completely modulated carrier, with high-level plate modulation. The design includes circuit for radio-frequency and modulator stages, selection of tubes (from tube manual), specification of circuit constants, and rough determination of such tube-operating conditions as expected output, d-c plate current, grid driving power, grid bias, etc. In this problem make full use of tube data given in tube manuals, and provide 100 per cent more driving power for each stage than estimates indicate will be needed.

9. Repeat Prob. 8, only using either control-grid or suppressor-grid modulation.

10. In a radio-telephone transmitter where the modulator does not have sufficient capacity to modulate the carrier completely, it is possible to adjust the linear amplifier following the modulator in such a manner as to make the degree of modulation of the output of the linear amplifier greater than the degree of modulation of the input wave. Explain how this could be done, and discuss the advantages and disadvantages of the arrangement.

11. In a radio-telephone transmitter, the effective Q of the tank circuits carrying the modulated waves determines the extent to which the higher side-band frequencies are discriminated against. If the total response for all the circuits is not to fall below 60 per cent for a modulation frequency of 5000 cycles in the transmitter of Fig. 290, what is the highest effective Q permissible at a carrier frequency of (*a*) 600 kc, (*b*) 1500 kc?

12. The discrimination against the higher audio frequencies mentioned in Prob. 11 is commonly compensated for by equalization in the audio-frequency system. Explain the limitations of such equalization when the degree of modulation is high at the higher modulation frequencies.

13. Draw a separate diagram of the negative feedback arrangement employed in the transmitter of Fig. 290, and explain the action in detail.

14. Draw a circuit showing a negative feedback arrangement in which the radio-frequency output of a linear amplifier is fed back to a preceding linear amplifier. Explain how the circuit functions and discuss its advantages and difficulties.

15. It will be noted in Figs. 288, 290, and 291 that the 50-kw transmitter is provided with a much more elaborate network for suppressing harmonics than the lower powered transmitters. Explain the reason for this.

16. Lay out a block diagram of a system for generating a single side-band signal having a carrier frequency of 20,000 kc, when the essential speech range of 250 to 3000 cycles is to be transmitted and it is assumed that filters are available which will effectively separate frequencies that differ by 1 per cent.

17. In a broadcast transmitter, 10 watts of energy radiated on an adjacent channel would produce appreciable interference. If this energy is assumed to be due to fourth-harmonic amplitude distortion at the transmitter when the carrier is completely modulated at 5000 cycles, calculate the maximum percentage distortion that is allowable for a 50-kw transmitter.

18. Explain why problems of adjacent channel interference become more important as the transmitter power increases.

19. What is the maximum permissible phase modulation (in degrees) that can be tolerated in a 50-kw broadcast transmitter if: (*a*) each second-order side-band frequency is to represent less than 10 watts of energy available for adjacent channel interference, (*b*) same for each third-order side-band component, (*c*) same for each fourth-order side-band component.

[In order to solve this problem it will be necessary to use a table of Bessel's functions, or to use the series expressions for the functions. In the latter case the computations can be simplified by noting that with the small arguments involved not more than one or two terms of the series need be used.]

20. Explain why the arrangement of Fig. 299 does not produce true phase modulation if the limiter is removed.

21. A linear amplifier is used to amplify the output of a controlled-carrier modulator. If the carrier amplitude for complete modulation is 500 watts, what will be the approximate plate loss and plate efficiency when (*a*) the full carrier is applied to the amplifier but there is no modulation at the moment, (*b*) the full carrier is completely modulated, (*c*) the carrier has half the full amplitude and there is no modulation, (*d*) the carrier is as in (*c*) but is fully modulated.

22. Assuming reasonable values of plate efficiency, estimate the total plate losses and the total power input (including modulator and radio-frequency tubes) required for a broadcast transmitter developing a 50-kw carrier when (*a*) unmodulated, and (*b*) completely modulated for the following systems: (1) high-level plate modulation using Class A audio system, (2) high-level plate modulation employing Class B audio system, (3) low-level modulation followed by conventional linear amplifier, (4) low-level modulation followed by Doherty high-efficiency linear amplifier, (5) outphasing system of modulation, (6) low-level modulation followed by linear amplifier employing dynamic shift of the operating point.

In making these estimates neglect all low-level radio-frequency stages and all modulation stages except the last stage of high-level systems.

23. Design a high-efficiency linear amplifier capable of delivering a carrier output of 10 kw. The design includes selection of tubes, circuit layout including specification of circuit constants, determination of tube operating voltages, and design of artificial lines.

CHAPTER XIII

RADIO RECEIVERS

107. Characteristics of Broadcast Receivers.—The most important characteristics of a receiver for radio-telephone signals are the sensitivity, the selectivity, and the fidelity. The sensitivity represents the ability of the receiver to respond to small radio-signal voltages, and is measured quantitatively in terms of the voltage that must be induced in the antenna by the radio signal to develop a standard output from the power amplifier. This standard output has been arbitrarily chosen as 0.05 watt in a non-inductive load resistance having a value corresponding to the load resistance into which the power amplifier is designed to operate. The sensitivity is arbitrarily defined as the effective value of the carrier voltage that must be induced in the antenna to develop this standard output when the carrier is modulated 30 per cent at a frequency of 400 cycles. The sensitivity is measured with the radio receiver tuned to give maximum response at the carrier frequency involved and with the volume controls adjusted for maximum volume. A curve showing the sensitivity of a typical broadcast receiver as a function of carrier frequency is shown in Fig. 307.

Selectivity is the property that enables a radio receiver to discriminate between radio signals of different carrier frequencies. Selectivity cannot be defined in a single term as can sensitivity but must be expressed in the form of curves, such as those of Fig. 307, which show the amount by which the signal input must be increased in order to maintain the standard output as the carrier frequency is varied from the frequency to which the receiver is tuned. These curves therefore indicate the extent to which interfering signals are discriminated against, and in general will depend somewhat on the carrier frequency.

Fidelity represents the extent to which the receiver reproduces the different modulation frequencies without frequency distortion. The fidelity of a radio receiver is expressed in curves, such as that of Fig. 307, which give the variation in audio-frequency output voltage as the modulation frequency of the signal is varied. In order to facilitate comparison, the output is expressed in terms of the ratio of actual output to the output obtained when the modulation frequency is 400 cycles.

Measurement of Receiver Characteristics.—The characteristics of a radio receiver are measured by using an artificial signal to represent the voltage

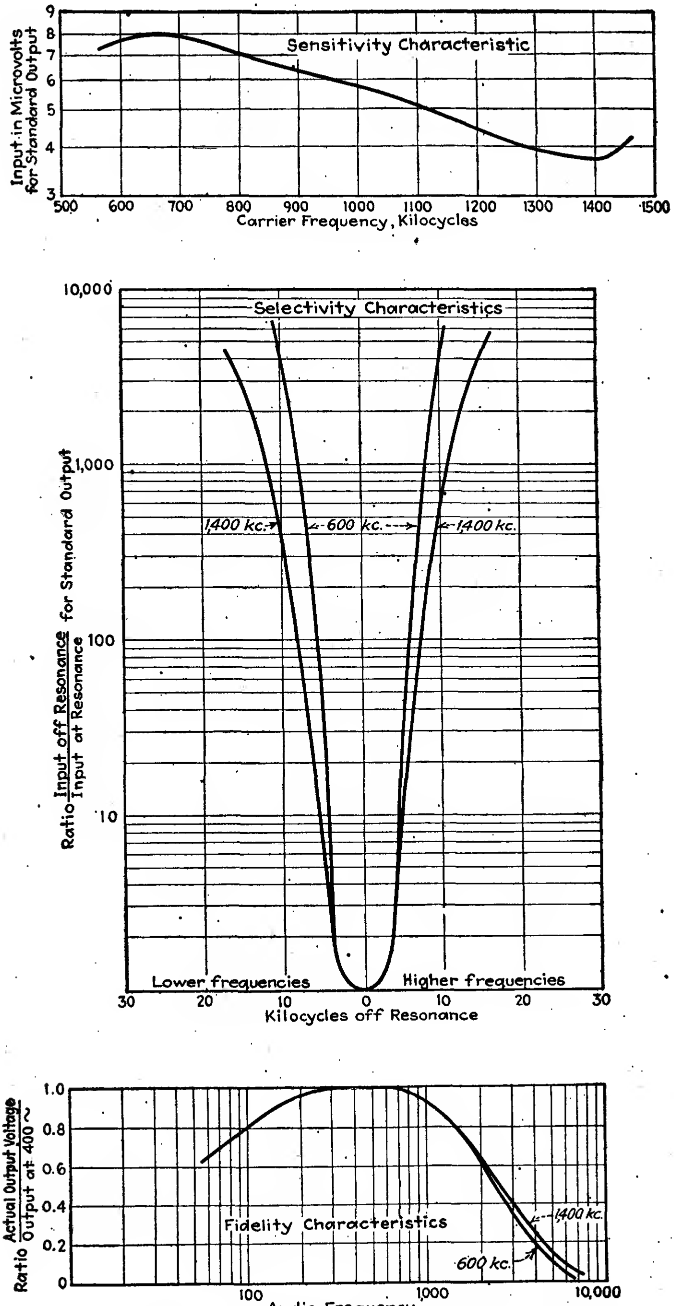


FIG. 307.—Typical sensitivity, selectivity, and fidelity curves of superheterodyne receiver.

that is induced in the receiving antenna. This artificial signal is applied to the input terminals through a network or "artificial antenna" that simulates the impedance of the actual antenna with which the receiver is to be used. In making this test the receiver output is determined by substituting a resistance load of the proper value for the loud-speaker and measuring the audio-frequency power in this resistance. The experimental set-up is illustrated in Fig. 308.

The equipment for producing the artificial signal is called a standard signal generator, and consists of a thoroughly shielded oscillator coupled to an attenuating system that is capable of producing known voltages from about 1 μv up to perhaps 200,000 μv. Provision is normally made for modulating this voltage any amount up to 100 per cent at all modulation frequencies that will be encountered.[1] The artificial antenna used in making receiver tests depends on the circumstances. For the regular broadcast band (550 to 1500 kc) the standard antenna consists of a capacity of 200 $\mu\mu$f, a self-inductance of 20 μh, and a resistance of 25 ohms, all in

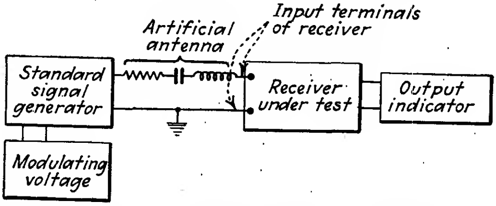


FIG. 308.—Schematic arrangement of equipment for making measurements of receiver performance.

series. For automobile radios the artificial antenna is a series capacity of 100 $\mu\mu$f combined with 60 $\mu\mu$f shunted across the receiver input terminals, while a resistance of 400 to 600 ohms is commonly employed in tests of short-wave performance of all-wave broadcast receivers.

Miscellaneous Characteristics and Considerations.—In addition to the sensitivity, selectivity, and fidelity, a broadcast receiver has other properties that warrant consideration. Among these are characteristics of the automatic-volume-control system, susceptibility to cross-talk interference, alternating-current hum level, noise level, audio power output obtainable without excessive distortion, and the nature and amount of the amplitude distortion in the output at different power levels. These characteristics are discussed further in Sec. 109, and can be readily found by appropriate methods.[2]

[1] The testing of broadcast receivers is of such importance that standards have been established for carrying on the more important types of measurements on radio receivers. These are described in greater detail in the Report of the Standards Committee of the Institute of Radio Engineers.

[2] For example, see F. E. Terman, "Measurements in Radio Engineering," 1st ed., Chap. IX, McGraw-Hill Book Company, Inc.

108. Typical Broadcast Receivers.—Practically all broadcast receivers are of the superheterodyne type, and accordingly have a schematic layout following that indicated in Fig. 309. The radio-frequency section is tuned to the signal frequency, and delivers a signal voltage at the grid of the first detector. The first detector (also called *mixer* and *converter*), together with its associated oscillator, converts the incoming oscillations to a fixed predetermined "intermediate" frequency by use of the heterodyne principle. The intermediate-frequency section between the two detectors is tuned to this predetermined difference frequency and delivers an intermediate-frequency voltage to the second detector where the modulation envelope is recovered from the wave by rectification. The resulting audio frequency is then amplified by the audio-frequency system and delivered to the loud-speaker for reproduction. While all superheterodyne receivers follow the general scheme outlined in Fig. 309, individual receivers differ greatly in detail. Thus the radio-frequency section may include one or more stages of amplification, or may be a simple tuned

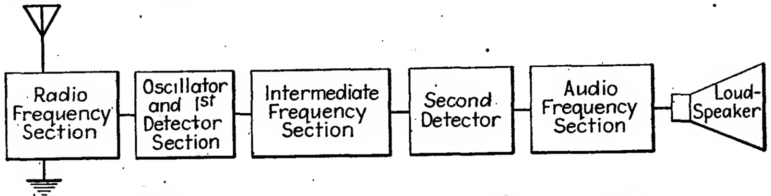


FIG. 309.—Schematic diagram of superheterodyne receiver.

circuit between the antenna and first detector. The first detector and oscillator section may involve any of the arrangements discussed in Sec. 88. Likewise, the intermediate-frequency section may consist of a simple tuned circuit between first and second detectors, or may include one or two stages of amplification. The second detector is nearly always of the diode type, while the audio-frequency system normally contains at least one stage of voltage amplification followed by a power amplifier commonly employing a push-pull connection.

All receivers also offer some combination of special features such as automatic volume control, manual volume control, tuning indicators, tone control, quieting arrangements, automatic frequency control, etc. The number and exact nature of these incorporated in a particular receiver varies greatly with the price and is affected by merchandizing considerations.

Typical Broadcast Receivers.—A general picture of the broadcast receiver situation can be obtained by considering a number of typical circuit arrangements. The receiver of Fig. 310 is a six-tube medium-price superheterodyne designed to cover the regular broadcast band (540 to 1800 kc) and one short-wave band (1800 to 6600 kc). In addition

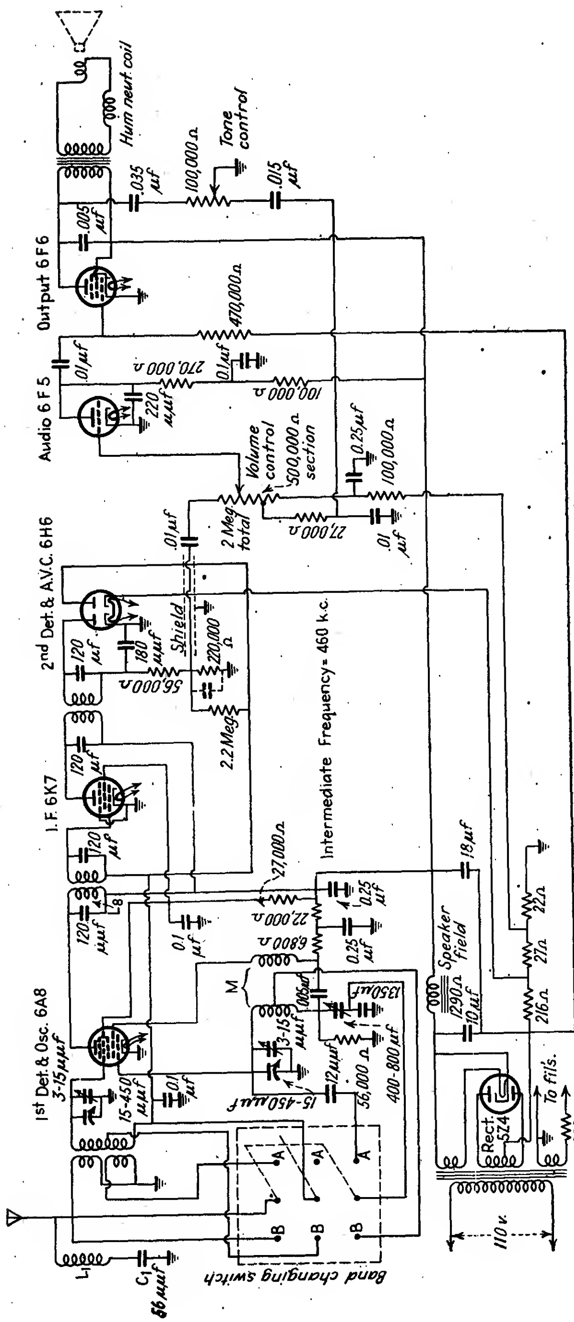


FIG. 310.—Circuit diagram of typical medium-priced superheterodyne receiver.

to the rectifier, there is a pentagrid converter, variable-mu intermediate-frequency amplifier, diode detector providing delayed automatic volume control, and an audio-frequency system consisting of a resistance-coupled high-mu voltage amplifier and a power pentode. A simple form of bass compensation is employed in connection with the manual volume control, and a tone control is also provided.

A very simple midget type of superheterodyne receiver employing only four tubes is shown in Fig. 311. The tube line-up, exclusive of the rectifier, consists of a pentagrid converter and a regenerative second detector of the grid-leak type, followed by a pentode output tube.

A relatively high priced all-wave broadcast receiver is shown in Fig. 312. This is a 13-tube receiver, with the frequency range of 540 to 18,000 kc divided into three bands. The tube line-up, not counting the

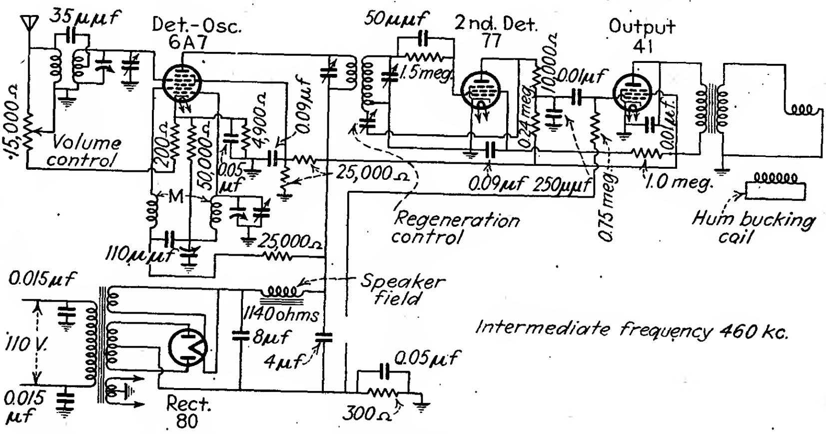


FIG. 311.—Circuit diagrams of typical midget superheterodyne of simplest and most inexpensive type.

two rectifier tubes, consists of one stage of tuned radio-frequency amplification, followed by a pentagrid converter, one stage of intermediate-frequency amplification, a combined diode detector and triode audio-frequency voltage amplifier in one envelope, a second audio-frequency voltage amplifier, and a pair of push-pull output tubes[1] capable of delivering approximately 20 watts of undistorted audio power. In addition there are three tubes employed in an automatic-frequency-control system, and one (essentially a lamp) for producing volume expansion. The receiver incorporates automatic volume control, manual volume control with bass compensation, and a tone control. The intermediate frequency transformers have three tuned circuits coupled to each other in

[1] These tubes are of the direct-coupled type described by Charles F. Stromeyer, General Theory and Application of Dynamic Coupling in Power Tube Design, *Proc. I.R.E.*, vol. 24, p. 1007, July, 1936.

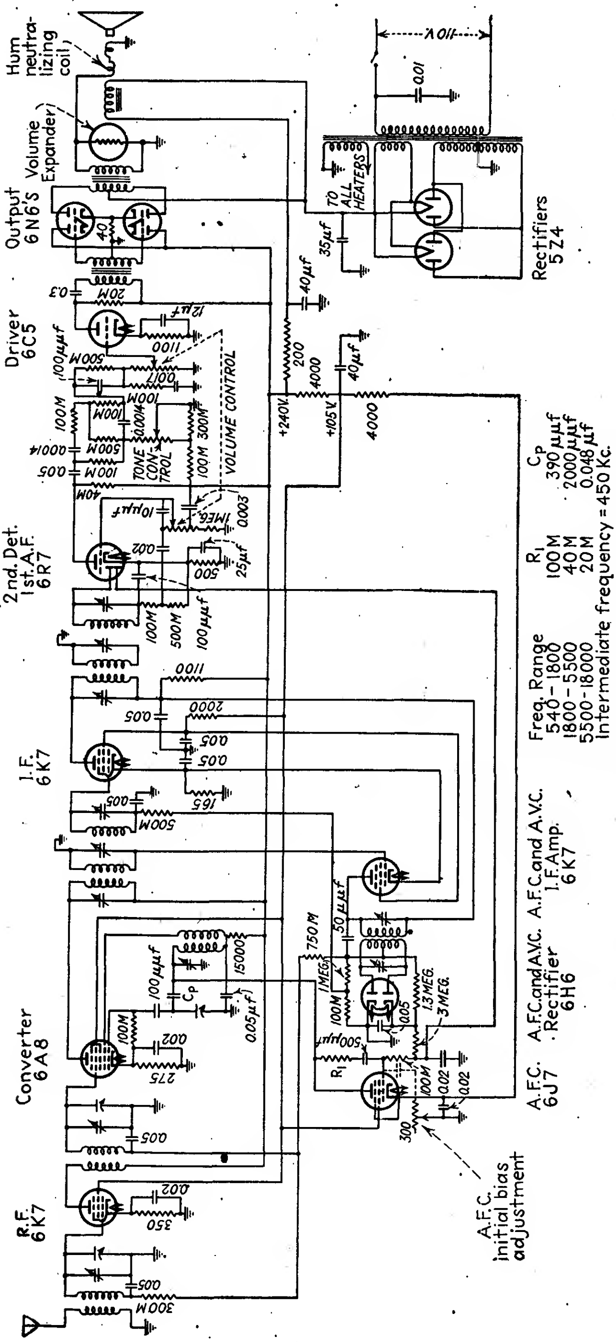


FIG. 312.—Simplified circuit diagram of 13-tube high-fidelity receiver featuring automatic tuning.

such a way as to produce a flat-top resonance curve about 10 kc wide and having very steep sides.

Another all-wave broadcast receiver incorporating a different combination of features is illustrated in Fig. 313. This is a 13-tube superheterodyne receiver, which covers the frequency range 540 to 23,000 kc in four tuning bands. In addition to the rectifier, the tube line-up consists of either one or two stages of tuned radio-frequency according to the frequency band, a first detector-oscillator combination consisting of a pentagrid converter tube with a separate oscillator feeding energy to the innermost grid of the tube, two stages of intermediate-frequency amplification, a diode-pentode tube functioning as the detector and first audio-frequency stage, a triode audio-frequency voltage amplifier, and a push-pull Class A power amplifier capable of developing 10 watts of undistorted power. The second detector uses separate diodes for automatic volume control (A.V.C.) and detection. An auxiliary tube is used to operate a tuning indicator, while the thirteenth tube is for "quieting" and functions only in the broadcast band. Two loud-speakers are provided, one for high and one for low and moderate audio frequencies. The band width to which the intermediate-frequency amplifier responds can be controlled by varying a resistance in series with a tertiary winding in the intermediate-frequency transformers. The manual volume control includes a bass-compensating arrangement. The tone control for reducing the high-frequency response is made inoperative when the selectivity on the intermediate-frequency amplifier is set for high fidelity.

Miscellaneous Types of Receivers.—There are several special-purpose types of broadcast receivers of sufficient importance to warrant consideration. These are the automobile radio, the battery radio, and the alternating-current–direct-current type of receiver. The automobile radio differs from the ordinary broadcast receiver primarily in that it obtains both filament and plate power from a 6-volt storage battery. This is accomplished by using 6-volt heater tubes and obtaining anode power by one of the systems described in Sec. 100. The circuits of the automobile receiver are essentially the same as any other receiver, except that the input is designed to operate with a different type of antenna.

In battery receivers economy of filament and anode power is the major consideration. Such receivers employ tubes especially designed for low filament power (such as the 2-volt series of tubes), and usually have Class B audio amplifiers in the power stage to reduce the anode power requirements.

The alternating-current–direct-current receiver is designed to operate either from 110 volts direct current or 110 volts alternating current. It employs heater-type tubes with 25-volt filaments that are connected in series across the 110-volt line. Plate power in the case of alternating-

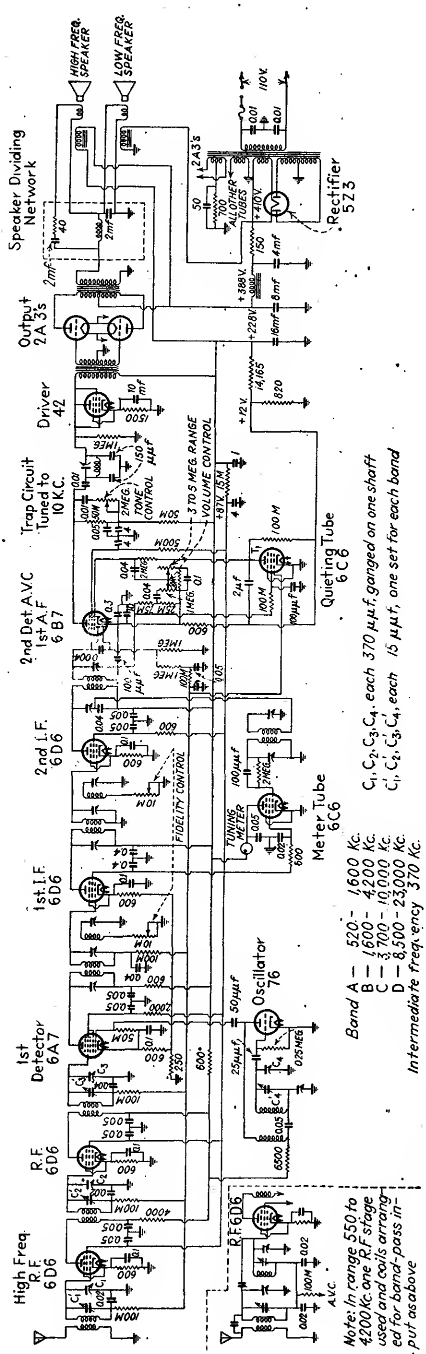


FIG. 313.—Simplified circuit diagram of 13-tube high-fidelity receiver featuring variable selectivity and a dual loud speaker system.

current supply is obtained by directly rectifying the line voltage without the use of a transformer, using a half-wave rectifier, while with direct-current lighting circuits the anode power is obtained directly without the need of rectifying, although for convenience the d-c current is usually passed through the rectifier tube and the filter system is used to eliminate ripple voltages. The general circuit arrangements employed in alternating-current–direct-current receivers are the same as in other receivers except that, since one side of the receiver is directly connected to the 110-volt power system, care must be taken in the construction to prevent shock hazard.

Some of the very cheap midget receivers employ a simple tuned-radio-frequency amplifier circuit. A typical receiver of this type is illustrated in Fig. 314, and has a tube line-up consisting of a rectifier, one stage of

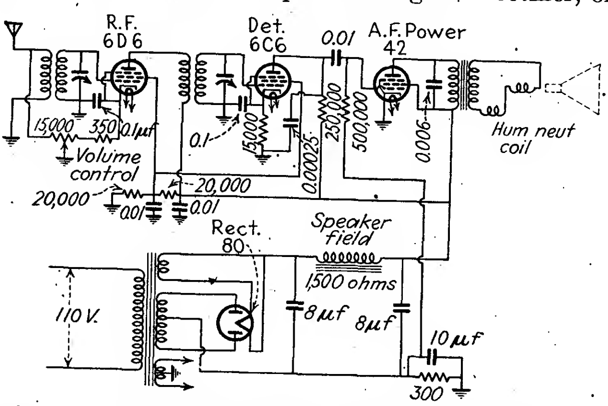


FIG. 314.—Simple tuned radio-frequency type of circuit such as employed in some of the very cheapest midget receivers.

tuned-radio-frequency amplification, a power detector, and a pentode power tube.

109. Broadcast Receivers; Miscellaneous Features.—Numerous special features and considerations are involved in broadcast receivers. Some of these have already been discussed in connection with amplifiers, detectors, power-supply systems, etc. The more important of the remainder are discussed below.

Automatic Volume Control.—All except the least expensive receivers are provided with some form of automatic volume control (abbreviated A.V.C.) to maintain the carrier voltage at the detector substantially constant. This is accomplished by biasing the grids of the radio-frequency, intermediate-frequency and sometimes the converter tubes negatively with a direct-current voltage derived by rectifying the carrier.[1] An increase in the signal hence increases the negative bias and

[1] In some cases less control bias is applied to the final intermediate-frequency amplifier than to the remaining tubes, in order to minimize distortion in this tube with strong signals. This is done in the receiver of Fig. 312.

thereby tends to counteract the increased signal by reducing the amplification, while, if the signal becomes weaker, the automatic-control bias is correspondingly less and the gain of the controlled tubes increases.

The actual details of automatic-volume-control systems may vary greatly. A simple diode detector can be used for both detection of the signal and for producing an automatic-volume-control bias by arrangements as illustrated in Fig. 239. In other cases a double diode detector is used, with one diode serving as the ordinary detector while the other develops the automatic-volume-control bias.

The performance of an automatic-volume-control system can be improved by delaying the control action until the signal voltage at the detector exceeds a predetermined value. Typical means of obtaining delay action are illustrated by the receivers of Figs. 310 and 313. In Fig. 310 the delay is obtained by connecting the anode of the second diode to the automatic-volume-control line and connecting the cathode of this anode to a point that is negative with respect to ground. With no signal the conductivity of this second anode causes the A.V.C. system to assume the potential of the second cathode, *i.e.*, a moderate negative bias. This bias is maintained on the A.V.C. system until the rectified signal develops a direct-current voltage equal to the second cathode voltage, at which point the second diode becomes non-conducting and the automatic-volume-control system functions in the usual manner. In the receiver of Fig. 313 separate diodes are used for detection and A.V.C., with the A.V.C. anode biased negatively by the amount of delay desired.

In some cases a separate intermediate-frequency amplifier stage is employed to supply voltage to the automatic-volume-control rectifier. When amplification of this type is employed to increase the voltage applied to the volume-control system, or when the rectified direct-current voltage is amplified before being used for control purposes, the system is termed *amplified automatic volume control.* By combining such amplified control with delay action, it is possible to maintain the carrier amplitude at the detector more nearly constant than is otherwise possible.

The automatic-volume-control bias is ordinarily applied to all radio-frequency and intermediate-frequency amplifier tubes (except possibly the tube preceding the detector) and also to the first detector. The controlled tubes are always of the variable-mu type in order to minimize cross talk.

Manual Volume Control.—All receivers are provided with a manually operated volume control to control the level of the reproduced sound. When the receiver also has automatic volume control, the manual control is practically always a potentiometer in the grid circuit of the first audio-frequency amplifier tube. In the case of very cheap receivers that have no automatic volume control, the manual control ordinarily operates by

varying the bias voltage on the grids of the radio-frequency tubes and is sometimes combined with an antenna shunting system that reduces the input voltage to the receiver at low volume settings.

Bass Compensation.—The characteristics of the human ear are such that, when sounds are reproduced at lower than normal volume levels, the low notes appear to be abnormally weak while, when the sound is reproduced at greater than normal level, the low notes appear to be abnormally loud.[1] In order to correct for this, the manual volume control of most receivers is arranged so that at low levels the intensity of the low notes is not reduced so much as is the volume of the higher pitched sounds. A simple example of such a bass-compensated volume control is found in the receiver of Fig. 310, where at low-volume settings the section of the potentiometer in use is shunted to ground through a resistance and capacity combination which has a lower impedance to high frequencies than to low frequencies and which thereby discriminates against the former. Other arrangements are used in the receivers of Figs. 312 and 313.

Tone Control.—Most receivers provide a *tone control* so that the listener may discriminate against the higher audio frequencies. The tone control is usually some form of resistance-capacity combination, with the arrangements shown in the receivers of Figs. 310, 312, and 313 being typical.

Tuning Indicators.—In receivers provided with automatic volume control, difficulty is encountered in tuning the receiver because the automatic-volume-control system tends to maintain the output constant even when not tuned exactly to resonance. At the same time, with slight mistuning the carrier is on the side of the response curve of the receiver, and considerable distortion results. Consequently, various devices have been developed for providing assistance in tuning the receiver.

The simplest form of tuning indicator is a direct-current meter through which flows the rectified output of the second detector. The tuning is then adjusted for a maximum deflection of the instrument. Another common arrangement employs a light valve actuated by the d-c current of the detector to vary the width of a shadow band. A more recent development in tuning indicators is a special miniature cathode-ray tube in which the luminous area is a sector of a circle having an angular spread determined by the negative bias applied to a pair of control electrodes. By deriving this bias from the automatic-volume-control system the size of the luminous area serves as an aid to tuning.

Automatic-frequency-control Systems.—The difficulty of accurately tuning a receiver having automatic-volume-control, as well as troubles caused by drifts in frequency of the beating oscillator when receiving short-wave signals, can be eliminated by an arrangement that will

[1] Further discussion of the characteristics of the ear is to be found in Sec. 148.

automatically shift the beating oscillator frequency so as to produce an intermediate frequency of exactly the proper value, provided the tuning is approximately correct. This is termed automatic frequency control (abbreviated A.F.C.) and is employed in the receiver of Fig. 312. Here the A.V.C. and A.F.C. systems are operated from a separate intermediate-frequency amplifier tube which has a load impedance consisting of a tuned primary circuit to which is coupled a tuned secondary circuit.[1] These are connected to a double diode tube as shown, so that the voltage applied to one diode is the vector sum of a component derived from the primary circuit and a component derived from the secondary circuit while the voltage applied to the second diode is the vector difference of these two components. The diodes are so arranged that two bias voltages are obtained, one proportional to the bias developed by one of the diodes and used for A.V.C. purposes, while the other is equal to the difference between the rectified voltages developed by the two diodes and is used to control the oscillator frequency. The operation of the device takes advantage of the fact that the voltage component produced by the secondary circuit is in quadrature with the voltage component obtained from the primary circuit provided the frequency of the signal is exactly in resonance with the secondary circuit. Under such conditions identical voltages are applied to the two diodes and the A.F.C. bias is zero. However, if the signal frequency differs from the resonant frequency of the secondary, there will be a phase shift away from the quadrature relation. This makes the voltage applied to one diode greater than the voltage applied to the other, so that the A.F.C. bias will be either positive or negative depending upon which diode receives the larger signal. This A.F.C. bias is applied to the grid of a pentode tube in which the plate circuit is in parallel with the tuned circuit of the oscillator, while the control grid is supplied with an exciting voltage that is 90° out of phase with the alternating voltage acting in the plate circuit. In such an arrangement the amplified grid voltage acting in the plate circuit draws from the oscillator a current 90° out of phase with the voltage across the oscillator circuit, so that the tube acts as a tuning reactance having a magnitude depending upon the amplification of the tube and hence upon the grid bias developed by the A.F.C. system. By proper arrangement of polarities, any deviation of the intermediate frequency from the proper value will cause the oscillator frequency to be shifted so as to reduce the deviation greatly.

Quieting Systems.—In tuning a sensitive receiver provided with automatic volume control, the noise output between stations is high because,

[1] A detailed discussion of the practical design of such A.F.C. systems is given by R. L. Freeman, Improvements in A.F.C. Circuits, *Electronics*, vol. 9, p. 20, November, 1936. D. E. Foster and S. W. Seeley, Automatic Tuning, Simplified Circuits, and Design Practice, *Proc. I.R.E.*, vol. 25, p. 289, March, 1937.

when no signal is being received, the A.V.C. system increases the sensitivity of the receiver to the maximum possible value. Arrangements for eliminating this interstation noise are variously known as Q circuits, quieting systems, squelch circuits, etc. A typical arrangement is employed in the receiver of Fig. 313, where a tube T_1 is so arranged that it biases the grid of the first audio tube beyond cut-off unless the grid bias on T_1 approaches or exceeds cut-off. By using the A.V.C. system to bias T_1 it is then possible to make the receiver inoperative until a carrier of predetermined amplitude is present.

Volume Expander.—In broadcasting, the operator at the transmitting station usually finds it necessary to reduce the intensity of the very loudest passages in order to prevent overmodulation of the transmitter, and also to increase the intensity of the weakest passages in order to prevent their loss in the noise level. The full volume range of the original signal can, however, be restored at the receiver by introducing an inverse effect, using an automatic volume expander such as discussed in Sec. 48. A very simple form of volume expander is used in the receiver of Fig. 312, and consists of a special lamp connected in shunt with the voice coil of the loud-speaker. During weak passages the filament of the lamp has sufficiently low resistance to shunt a considerable portion of the output power around the loud-speaker, but during loud passages the lamp filament becomes hot, thereby increasing the filament resistance and so increasing the proportion of the total output that is delivered to the loud-speaker.

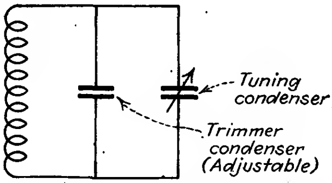


FIG. 315.—Trimmer system for tuned radio-frequency amplifier circuit.

110. Alignment.—In the superheterodyne receiver the radio-frequency circuits are tuned to the incoming signal being received while the resonant frequency of the oscillator circuit must at the same time differ from the signal frequency by an amount equal to the intermediate frequency. This introduces a problem of alignment, or tracking, since in the ordinary receiver the various circuits are adjusted simultaneously by a single control.

The usual procedure is to use tuning condensers in which the various sections are as nearly identical as possible, and then to attain the necessary tracking by the use of trimmer condensers and by proper coil inductances. In a receiver covering only the regular broadcast band the radio-frequency stages are aligned at the high-frequency end of the band by means of adjustable shunt padding condensers, as illustrated in Fig. 315, while at the low-frequency end of the band exact alignment is obtained by bending the end plates of the condensers. When care is taken to insure that the coils and condensers are initially very nearly

alike, satisfactory tracking will then be obtained over the entire band, except possibly for the antenna circuit, where the wide variety of antenna constants encountered makes perfect alignment under all conditions impossible.

The oscillator of a superheterodyne receiver is ordinarily tuned by a condenser that is identical with the condenser gangs used with the radio-frequency circuits, and is made to track, (*i.e.*, to produce the required difference frequency) by using series and parallel trimmer condensers

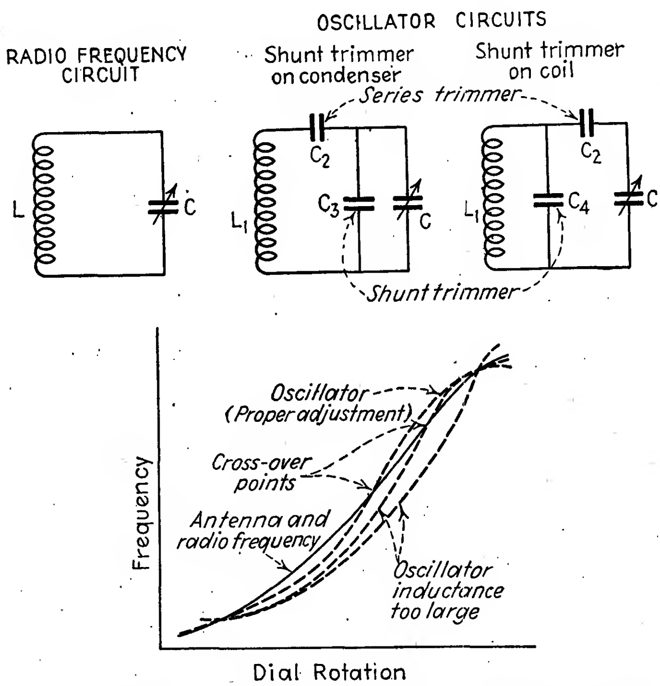


FIG. 316.—Trimmer systems for superheterodyne oscillator circuits, together with schematic curves showing kind of tracking obtained for typical cases.

as shown in Fig. 316, together with the proper choice of tuning inductance. Analysis of the equivalent circuits of Fig. 316 shows that by a suitable choice of C_2, C_3 or C_4, and L_1/L it is possible to make the difference in the resonant frequency of the radio-frequency and oscillator circuits exactly equal to an assigned value at three settings of the tuning condenser. Furthermore, if two of these frequencies of perfect tracking are near the extremes of the tuning range, while the third is located near the middle, then the actual tracking will be almost perfect throughout the tuning range, as shown in Fig. 316. A mathematical analysis of the problem involves a rather complicated series of manipulations based upon the

circuits of Fig. 316, but the essential results are incorporated in the following equations:[1]

For shunt trimmer on condenser:

$$\left.\begin{aligned} C_2 &= C_0 f_0^2\left(\frac{1}{n^2} - \frac{1}{l^2}\right) \\ C_3 &= \frac{C_0 f_0^2}{l^2} \\ L_1 &= L\left(\frac{l}{m}\right)^2\left(\frac{C_2 + C_3}{C_2}\right) \end{aligned}\right\} \qquad (204)$$

For shunt trimmer on coil:

$$\left.\begin{aligned} C_2 &= \frac{C_0 f_0^2}{n^2} \\ C_4 &= \frac{C_0 f_0^2}{(l^2 - n^2)} \\ L_1 &= L\left(\frac{l}{m}\right)^2\left(\frac{C_2}{C_2 + C_4}\right) \end{aligned}\right\} \qquad (205)$$

where, in addition to the notation of Fig. 316,

f_0 = intermediate frequency
f_1, f_2, f_3 = frequencies at which exact tracking is required
$a = f_1 + f_2 + f_3$
$b^2 = f_1f_2 + f_1f_3 + f_2f_3$
$c^3 = f_1f_2f_3$
$d = a + 2f_0$
$l^2 = (b^2d - c^3)/2f_0$
$m^2 = l^2 + f_0^2 + ad - b^2$
$n^2 = (c^3d + f_0^2l^2)/m^2$
$C_0 = 25{,}330/Lf_0^2$ = capacity required to tune L to f_0.

All frequencies are expressed in megacycles, inductance in microhenries, and capacities in micromicrofarads. These formulas assume that the distributed capacity of the coil is a part of the tuning capacity C, which though an approximation is still sufficiently accurate for ordinary design purposes.[2]

[1] This follows Laboratory Series Report UL-8, RCA Radiotron Company. Other equivalent formulas are given by V. D. Landon and E. A. Sveen, A Solution of the Superheterodyne Tracking Problem, *Electronics*, vol. 5, p. 250, August, 1932; Hans Roder, Oscillator Padding, *Radio Eng.*, vol. 15, p. 7, March, 1935; A. L. M. Somerby, Ganging the Tuning Controls of a Superheterodyne Receiver, *Wireless Eng. and Exp. Wireless*, vol. 9, p. 70, February, 1932.

[2] In actual practice the effect of the coil capacity is taken into account by a slight experimental readjustment of the shunt-trimmer capacity during the alignment process.

Tracking Problem in All-wave Receivers.—In all-wave receivers the tracking problem is complicated by the fact that it is not possible to bend the end plates of a condenser gang to line up the low-frequency end of one band without thereby throwing out the adjustment for all other bands. The usual arrangement in all-wave receivers is to employ separate shunt trimmers associated with each coil. In the radio-frequency section, the high-frequency end of each band is aligned by adjustment of these trimming condensers, while accurate control of the coil inductance, maximum condenser capacity, stray capacity, etc., is ordinarily depended upon to insure proper lining up at the low-frequency end of each tuning range.[1]

Each oscillator coil of an all-wave receiver is provided with its own series and shunt-trimming condensers following the arrangement of Fig. 316. This makes it possible to adjust each band separately. The shunt trimmer is always adjustable, while an adjustable series trimmer is used for the lower frequency bands. In the high-frequency bands the required series capacity is so large that the most practical arrangement is a fixed condenser. Tracking at the low-frequency end of the band is then obtained either by careful manufacturing control of the coil and condenser constants to insure uniformity or by individually adjusting the inductance of the oscillator coil.

In some of the less expensive two-band receivers only one set of coils is provided, and band changing is accomplished by short-circuiting a portion of each coil, as in the receiver of Fig. 310. In such an arrangement the receiver is aligned only for the regular broadcast band, and manufacturing control of coil and condenser uniformity is depended upon to maintain at least passable tracking for the short-wave band.

Experimental Procedure for Aligning Receivers.—The alignment of a receiver can be readily carried out experimentally by using a test oscillator and some form of output indicator. The first step is to line up the intermediate-frequency amplifier. This can best be done with a frequency-modulated test oscillator combined with an output indicator consisting of a cathode-ray tube and a synchronized sweep circuit,[2] but a simple test oscillator with ordinary output indicator can be used if nothing better is available. The alignment is carried out by working step by step through the intermediate-frequency circuits from second detector toward the first detector, always applying the test oscillator to the grid of the tube immediately preceding the circuits under adjustment.

[1] In some cases, particularly in the high-frequency bands where a small change in the inductance of the lead wires represents a considerable percentage change in the total inductance of the tuned circuit, it is necessary or desirable to adjust the circuit inductance individually for each coil by varying the position of a turn of the coil or by adjustment of the lead wires.

[2] See F. E. Terman, *op. cit.*, p. 333, for further discussion of such test devices.

The next step is to align the radio-frequency and oscillator circuits for the broadcast band. The receiver dial and test oscillator are set to a frequency somewhere near the high-frequency end of the band, and the test oscillator is connected to the antenna input and adjusted so that only a small output is obtained. The shunt-trimming condensers on the radio-frequency stages are then adjusted until the output is maximum, after which the shunt trimmer on the oscillator condenser is varied to give maximum receiver output.[1] The receiver dial and test oscillator are then set to a frequency at the low end of the band, and the series padding condenser of the oscillator is adjusted for maximum response. If the receiver covers only the regular broadcast band, it is also permissible to improve the alignment of the radio-frequency stages by bending the end plates of the condensers. Finally, receiver and test oscillator are retuned to the original high frequency, and the shunt-trimmer condenser on the oscillator is checked to make sure that the change in the series padding condenser has not affected the required shunting condenser. The receiver is now perfectly aligned at the high-frequency end of the band, and the oscillator tracks at both high- and low-frequency ends and also at an in-between frequency determined by the oscillator inductance which the manufacturer placed in the receiver.

In all-wave receivers this process is repeated for each band, except that for the higher frequency bands the series trimmer on the oscillator is usually non-adjustable.

In the design of receiver tuning systems the usual procedure consists in calculating the proper oscillator inductance by Eq. (204) or (205), and then with this inductance in the circuit aligning the high- and low-frequency ends of the band by series and shunt trimmers as explained. As a final check, the exact location of the third frequency of perfect tracking (the "cross-over" point of Fig. 316) is obtained experimentally by noting the extent to which one of the trimmers must be readjusted to obtain perfect alignment at different dial settings. If the crossover point is at too high a frequency, the oscillator inductance is then too high, and vice versa, as indicated in Fig. 316.

111. Broadcast Receivers; Construction and Design Considerations. Broadcast receivers are mounted on a chassis of plated sheet iron bent in the form of an inverted tray and suitably punched to receive sockets, coils, transformers, etc., which are usually held in place either by eyeletting or by self-tapping screws. This chassis is then slipped into a cabinet which is essentially a piece of furniture rather than an electrical device, and is connected to the loud-speaker by means of a flexible cord terminated in a

[1] All oscillator adjustments are preferably made while rocking the receiver tuning dial slightly in order to take care of any slight interaction between the oscillator and radio-frequency stages.

plug. Photographs showing the construction of typical receivers are shown in Figs. 317 to 321. The tubes, transformers, coil assemblies,

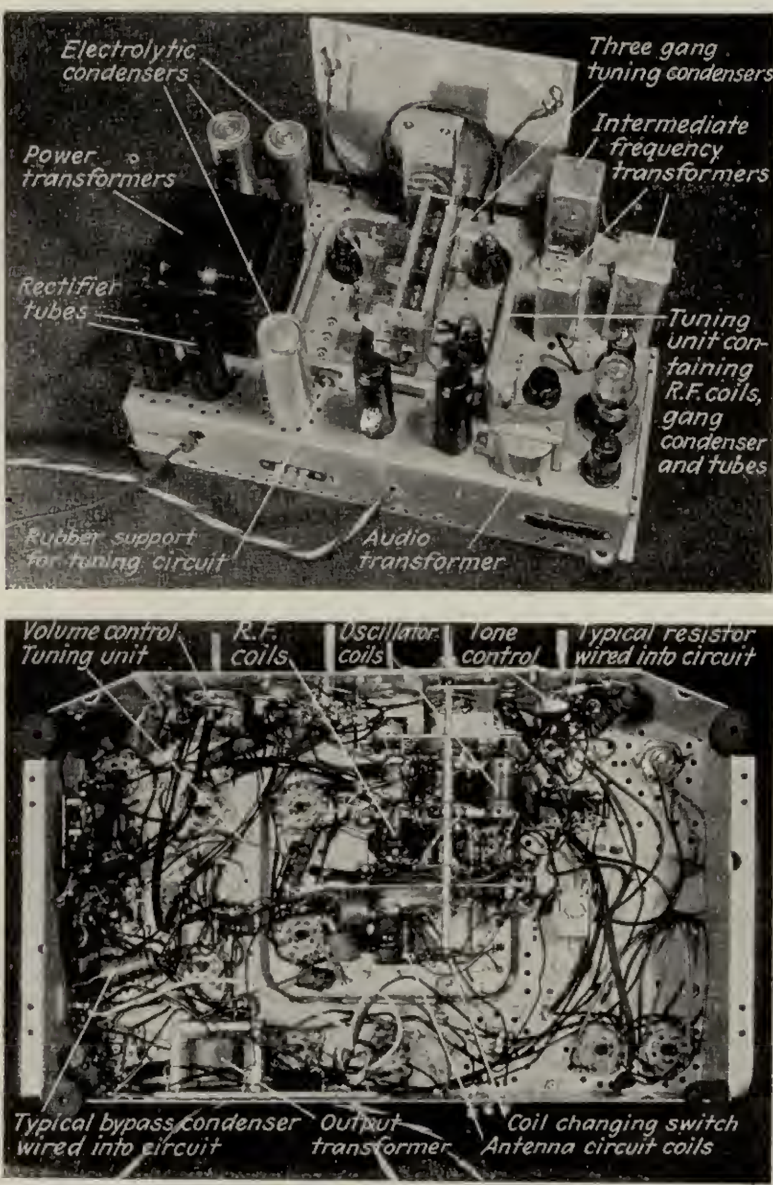


FIG. 317.—Top and bottom of chassis of high-fidelity receiver of Fig. 312.

gang tuning condenser, electrolytic condensers, and other bulky items are usually mounted on the top side of the chassis, while small parts such as

by-pass condensers, resistors, etc., are placed below along with most of the wiring. The arrangement of parts is carefully worked out to avoid troubles from hum, regeneration, etc., while at the same time providing for accessibility and ease of construction. In the wiring, leads carrying high-frequency currents, and also low-level audio-frequency leads, are as short and direct as possible. Shielded wires are sometimes employed and it is always necessary to place the critical leads carefully in order to avoid regeneration and whistles. The exact arrangement of the other leads is unimportant, and these are frequently run "haywire," as shown in Fig. 317, in order to reduce the cost of production.

The maximum possible use is made of sub-assemblies in the construction, because this simplifies the production and makes it easier to locate trouble. Thus coil assemblies with their associated shield cans, trimmer condensers, etc., are commonly constructed and tested separately and mounted upon the chassis as a single unit.

Cost is of prime importance in the design and construction of commercial equipment. Carbon-stick resistors with pigtail leads and tubular by-pass condensers with cardboard covering and pigtail leads are employed wherever possible. These are wired directly into the circuits so as to be self-supporting or are arranged on mounting strips, according to the circumstances. Chokes, coils, and transformers are avoided wherever possible in order to reduce cost. The filter system for supplying the voltage to the power stage is nearly always a single-section shunt-condenser arrangement employing electrolytic condensers together with a choke provided by the field of the loud-speaker. Additional filtering for the remaining tubes is obtained by resistance-condenser combinations as seen in Figs. 310 to 314.

The radio-frequency coils of broadcast receivers are very small, and are usually, although not always, mounted in a shield can. The tuning coils for the regular broadcast band are either single layer or bank wound, preferably of Litz wire, while coils for the high-frequency bands of all-wave receivers are nearly always a single layer of solid wire. Uniformity in coil inductance is obtained by grinding the coil forms to exact dimensions, and is aided by winding the turns in threads. Residual variations occurring in production can be taken care of by classifying individual coils according to their exact inductances, or by individual adjustment of the position of the end turns.

Intermediate-frequency coils are commonly arranged in pairs with the primaries and secondaries separately tuned to provide a band-pass effect, with the coupling usually slightly more than the critical value.[1] Univer-

[1] It is to be noted that, when the output of an intermediate-frequency amplifier is applied to a diode detector, the input resistance of the diode reduces the effective Q of the secondary circuit. This reduces the selectivity and gain, and makes it neces-

sal-wound coils are always employed, and some form of iron-dust core is very common. Adjustment of the resonant frequency of the tuned circuits is obtained either by an adjustable condenser or, in the case of some iron-core types, by a screw-driver adjustment of the core position.

Coil and Band-switching Arrangements for All-wave Receivers.—In all-wave receivers the switching arrangement for changing coils must have low resistance, permit short leads, and introduce negligible coupling to

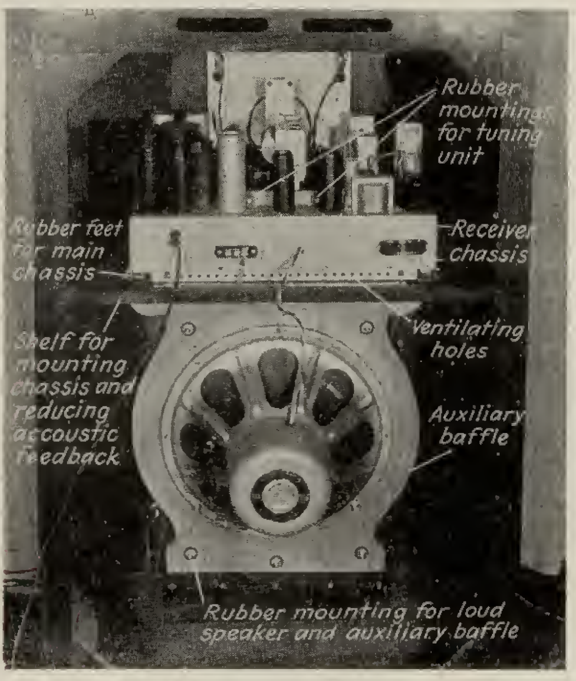


FIG. 318.—Arrangement of chassis and loud-speaker in cabinet, showing means used to minimize acoustic feedback (same chassis as in Fig. 317).

other circuits. The physical arrangement varies with the manufacturer, but the tendency is toward placing tuning condensers, radio-frequency and oscillator coils, and radio-frequency, oscillator, and first detector tubes, together with the band-selecting switch, in a sub-assembly which can be inserted in the chassis as an independent unit. Such construction is employed in the receivers of Figs. 317 and 321. This arrangement facilitates trouble shooting, simplifies production, and makes for flexibility in manufacture by allowing the same tuning unit to be placed in different models of receivers.

---

sary to use closer coupling between primary and secondary coils than when the output is applied to the grid of an amplifier tube.

A variety of coil constructions is employed on commercial all-wave broadcast receivers. In some cases separate coil forms are used for each band, and are mounted directly on the switch as shown in Fig. 317, while

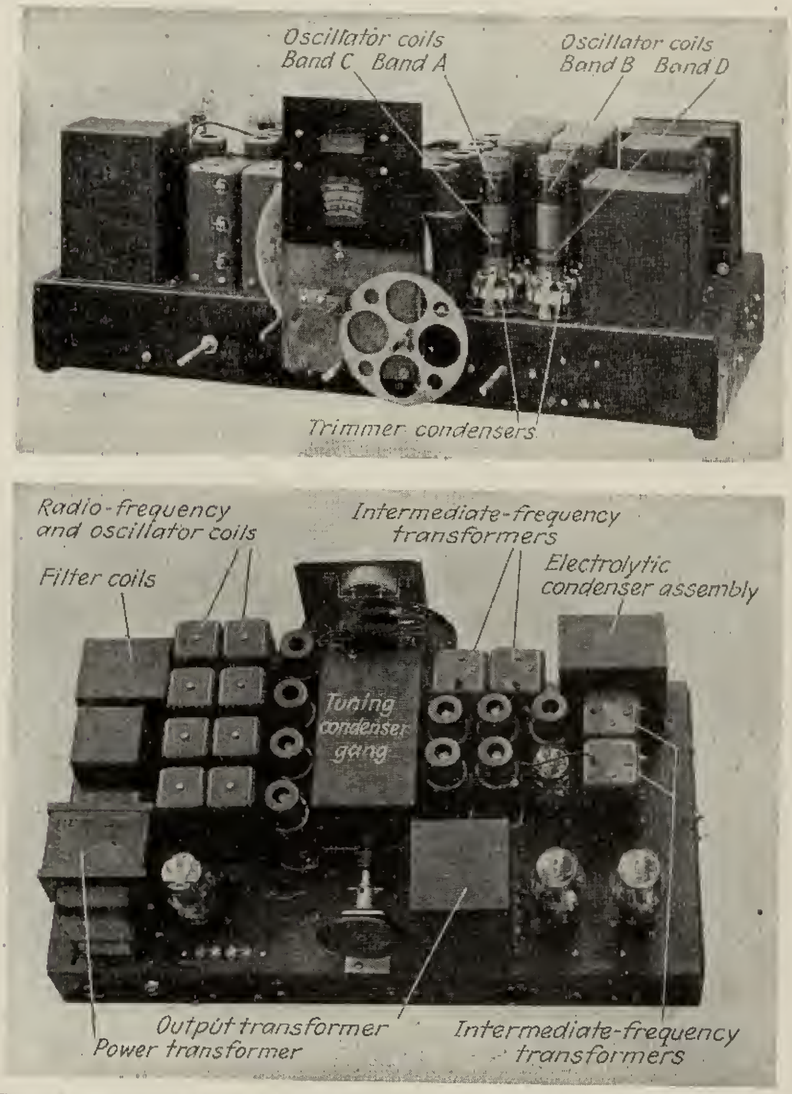


FIG. 319.—Chassis of receiver of Fig. 313. In the lower view the shield cans are removed from the oscillator coil assemblies. Note the two sets of coils in the same shield, and the individual trimmer condensers at the base of the coil form.

in other instances several coils are mounted on the same form as illustrated in Fig. 319, with wires brought out from the various sections to the switch. Coils are usually placed in shield cans, although they can be left

in the open, as in Fig. 317, if carefully placed. The unused coils, particularly the ones for the lower frequency bands, are short-circuited to prevent their acting as resonant secondary circuits tuned by distributed

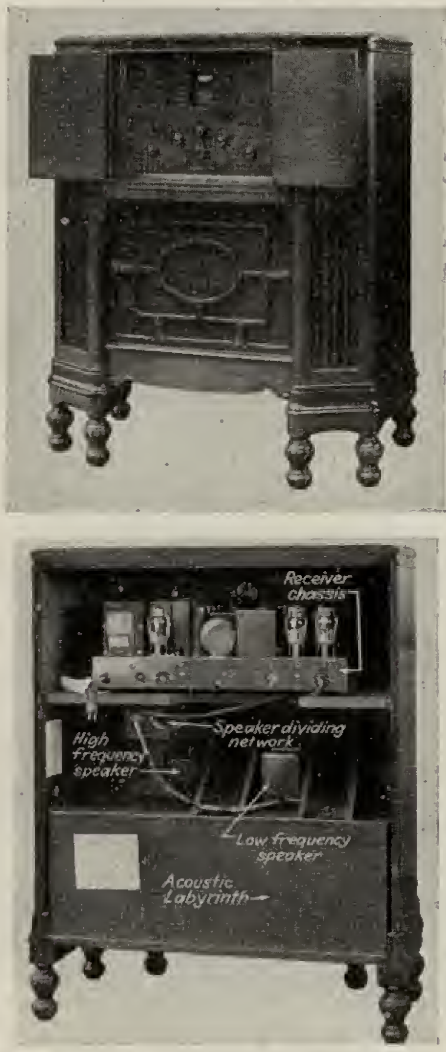


FIG. 320.—Arrangement of chassis and loud-speakers in the receiver of Figs. 313 and 319. Note the two loud-speakers and the acoustical labyrinth.

capacities. Tapped coils are used only in the less expensive two-band receivers, such as in the receiver of Fig. 310.

Spurious Responses in Superheterodyne Receivers.—Superheterodyne receivers when not carefully designed are troubled with a variety of spurious responses which appear as whistles. The most important of

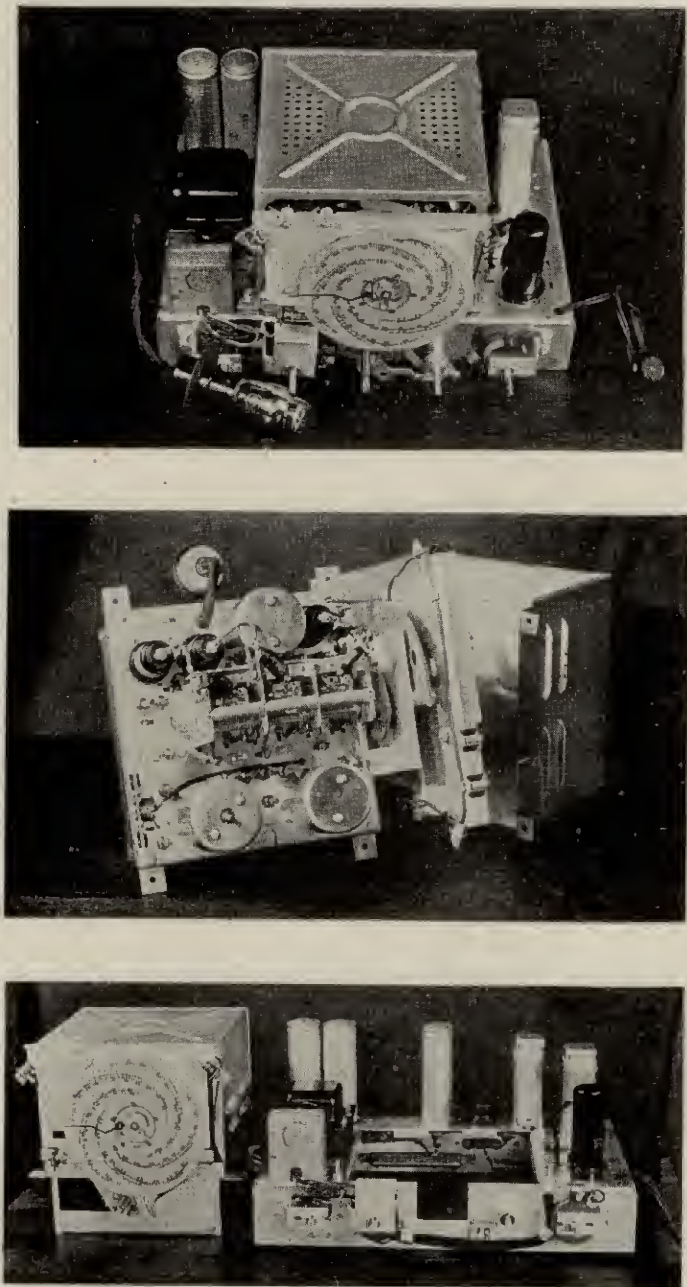


FIG. 321.—Chassis of an all-wave receiver in which the radio-frequency and oscillator circuits, together with associated tubes, are mounted in a separate unit that can be removed from the chassis.

these arise from image-frequency signals, from signals of intermediate frequency, and from harmonics of the intermediate frequency.[1]

Image response arises from the fact that there are two possible signal frequencies that will form the required intermediate frequency at any dial setting. Thus with an intermediate frequency of 50 kc, such as was employed in the early superheterodynes, the local oscillator would be adjusted to 1050 kc to receive a 1000-kc signal. Another signal of 1100 kc would, however, also produce a difference frequency equal to the intermediate frequency, and so might be simultaneously heard in the receiver output. The only remedy for this is to prevent the undesired image signal from reaching the grid of the converter tube. This requires that the radio-frequency section have an adequate number of tuned circuits and that the intermediate frequency be high enough to enable these circuits to suppress the image frequency. In receivers covering only the regular broadcast band the intermediate frequency is commonly about

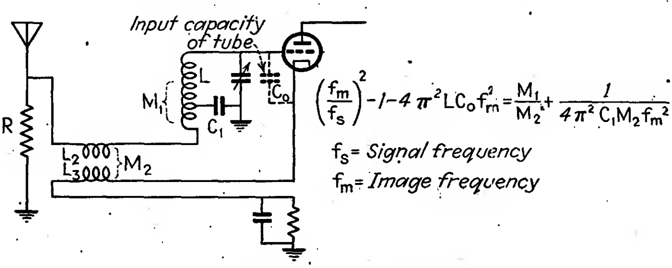


FIG. 322.—Typical circuit for suppressing image signals together with equations for maximum suppression.

262.5 kc, which makes the image frequency differ from the desired frequency by 525 kc. In all-wave receivers a higher intermediate frequency is employed, usually about 460 kc, and even then the image suppression is none too good at the highest frequencies, where the 920-kc difference between desired and image frequencies represents only a small percentage of the desired carrier frequency.

The suppression of the image response can be greatly improved by means of special suppressing circuits, of which the arrangement of Fig. 322 is typical.[2] Here the antenna is brought to ground through a coupling coil L_2, a portion of the tuning conductance L, and a capacity C_1 of the order of ten times the maximum tuning capacity. The resistance R is for broadening the antenna resonance. The effect of the image voltage

[1] An extensive discussion of all the possible sources of spurious responses in a superheterodyne receiver is given by Howard K. Morgan, Interfering Responses in Superheterodynes, *Proc. I.R.E.*, vol. 23, p. 1164, October, 1935.

[2] For further discussion of this and other circuits for suppressing image signals, see Harold A. Wheeler, Image Suppression in Superheterodyne Receivers, *Proc. I.R.E.*, vol. 23, p. 569, June, 1935.

produced at the grid of the tube is largely neutralized by the voltage that the antenna current induces in the cathode coil L_3. With proper circuit proportions the suppression can be made almost complete at two carrier frequencies and will be high at frequencies in between. At the same time, the desired carrier is not balanced out. The necessary design formulas are given in Fig. 322, and the suppression obtained is independent of the antenna constants.

The importance of image suppression is greatest in the regular broadcast band where the stations are numerous, powerful, and spaced at narrow frequency intervals. The image suppression of ordinary all-wave receivers in the higher frequency bands is rather poor, but fortunately the higher frequency stations are not so closely spaced in the frequency spectrum so that there is less likelihood of an image-frequency signal being present.

Strong signals of intermediate frequency from ships, etc., may reach the grid of the first detector when insufficient preselectivity is present, and will then be amplified by the intermediate-frequency amplifier. Interference of this type can be readily eliminated by designing the antenna input circuit in such a manner as to discriminate against signals of intermediate frequency. Thus in the receiver of Fig. 310 inductance L_1 and condenser C_1 are in series resonance at the intermediate frequency and so by-pass signals of intermediate frequency directly to ground.

When the intermediate-frequency harmonics that are produced by the second detector get coupled into the radio-frequency input circuits, interference in the form of whistles can be expected at certain places on the dial. Thus, if the intermediate frequency is 452 kc and the signal being received is 900 kc, the second harmonic of the intermediate frequency is 904 kc and, if this gets back into the input circuits, it will give a 4000-cycle beat note with the desired signal. The remedy for trouble of this sort is adequate by-passing of the circuits of the second detector, together with shielding and proper placement of circuit elements.

Cross-talk.—Cross-talk is the name given to interference resulting from interaction of radio signals, and it can be divided into two principal types. The first kind of cross-talk is produced by heterodyne detection of two signals having a frequency difference lying within the tuning range of the receiver. For example, when one broadcast station is operating on 1400 kc and another on 600 kc, heterodyne detection of the two carrier waves will result in the production of an 800-kc difference frequency, which will be heard when the receiver is tuned at or near this difference frequency. The production of such cross-talk requires that the two interfering signals reach the grid of the first radio-frequency tube and that the characteristic curve of this tube be curved at the operating point. The magnitude of the cross-talk output can be determined by consider-

ing the first radio-frequency tube to be a weak-signal detector of the anode type. The results of such an analysis show that the magnitude of the cross-talk is proportional to the product of the interfering signal voltages that reach the grid, and increases with the curvature of the tube characteristic. Heterodyne cross-talk is therefore most prominent when the interfering signals are produced by powerful local stations and when the automatic volume control places the operating point of the radio-frequency tubes very close to cut-off. This type of cross-talk can be suppressed almost completely by using a tuned input circuit between the grid of the first tube and the antenna, since such a circuit will prevent at least one of the interfering signals from reaching the grid of the first tube.

The second kind of cross-talk is heard under the following circumstances: The receiver is tuned to a powerful local station—the "desired" signal—which is so strong as to require a low-gain condition of the automatic volume control. At the same time there is another powerful local station—the "unwanted" signal—operating on a frequency not greatly different from that of the station being received. During the interval in which the desired station is sending out an unmodulated carrier wave the modulation of the unwanted signal will be heard, but if the desired station ceases to radiate its carrier wave the interfering signals from the unwanted station disappear. Such cross-talk is caused by the unwanted signal modulating the carrier wave of the desired signal by the mechanism discussed in Sec. 55. This is much more troublesome than the first kind of cross-talk because it can occur when the frequencies of the unwanted and desired signals are only slightly different.

This type of cross-talk is a result of third-order curvature in the tube characteristic and can be evaluated in terms of the cross-talk factor discussed in connection with Eq. (137). Such cross-talk was very common in the early radio receivers, but has been largely eliminated by the development of the variable-mu tube with its low third-order curvature at low plate currents.

Alternating-current Hum.—An alternating-current power-line hum often appears in the output of radio receivers energized by alternating current. The chief causes of such hum are ripple in the output of the rectifier-filter system, hum pick-up by the input circuits of the first audio-frequency tube, and power-line disturbances that are coupled into the receiver through the power transformer. Hum from rectifier ripple is particularly troublesome when a push-pull power amplifier is not used, and in such cases it can be greatly reduced by providing the filter choke with a hum-bucking coil that is connected in series with the voice coil of the loud-speaker as in Fig. 310. By suitable design the ripple voltage induced in this bucking coil will neutralize the hum voltage in the amplifier output. Hum pick-up by the input circuits to the audio-frequency

system can be prevented by shielding the grid lead as in Fig. 310 or by careful placement with respect to wires carrying a-c currents. Power-line disturbances represent radio-frequency oscillations produced by spark discharges and failing insulation; they can be eliminated either by the use of a grounded electrostatic shield between the primary and secondary of the power transformer or by by-pass condensers from the 110-volt leads to ground, or both.

Noise.—When a radio receiver is adjusted for high sensitivity, there is always a background of noise in the form of hiss, crackles, etc. A portion of this arises from thermal agitation and shot-effect noises developed within the receiver itself as explained in Sec. 49, while the remainder represents radio-frequency oscillations produced either by such natural causes as lightning, *i.e.*, static, or by such man-made devices as electrical appliances, automobile ignition systems, etc.

Thermal agitation and shot effect produce the hiss heard in home receivers when the antenna is disconnected, and they set an absolute limit to the maximum sensitivity that can be utilized. When a receiver is properly designed, this noise should come primarily from the thermal agitation in the input circuit of the receiver, rather than from shot effect in the first tube or thermal agitation in subsequent tuned circuits.[1] The extent to which this ideal is realized can be tested by disconnecting the antenna and independently tuning the input circuit through resonance (or, when this is impractical, by short-circuiting the input tuned circuit) and noting whether there is much change in the hiss noise.[2]

Static is discussed further in Sec. 127 and does not influence the design of broadcast receivers since there is little that can be done in such receivers to reduce the effects of static.

The man-made noise level is highest in thickly settled regions, particularly industrialized districts where a large amount of electrical equipment is in operation. The general level of such disturbances can be kept down to some extent by attention to the more important noise sources, but at best the residual noise level is much higher in urban districts than in rural areas.

The noise problem is particularly difficult to handle when a radio receiver is operated in close proximity to an internal-combustion engine. In the case of automobile radios operating in the regular broadcast and

[1] The local oscillator associated with the first detector of a superheterodyne receiver very often introduces a considerable amount of noise. The amount varies with the oscillator tube, and can in any case be made negligible compared with thermal agitation in the input circuit by employing an efficient coupling system between antenna and first detector, or better yet by the use of a stage of radio-frequency amplification.

[2] See F. B. Llewellyn, A Rapid Method of Estimating the Signal-to-noise Ratio of a High Gain Receiver, *Proc. I.R.E.*, vol. 19, p. 416, March, 1931.

police bands, the methods of suppressing ignition noise have become rather well standardized. They include such things as by-pass condensers around the generator, radio-frequency filters in unshielded wires going to the dome light, suppressor resistances in series with the spark plugs to damp out the oscillations that would otherwise be produced, etc. It is also sometimes necessary to bond certain parts to the frame in order to prevent noises from intermittent contacts. The best combination to use varies with the model and make of car and is usually worked out by the car manufacturer. The automobile radios themselves are completely inclosed in a metal box to provide electrostatic shielding, and a filter consisting of by-pass condensers and radio-frequency chokes, as shown in Fig. 278, is used to prevent noise from entering through the battery leads. Most receivers also have a low-pass filter in the antenna input circuit which passes the desired signal frequencies but prevents noise voltages of very high frequency from reaching the grid of the first tube and introducing cross-modulation.

Ignition noise is particularly intense at the higher frequencies, and is correspondingly difficult to eliminate in short-wave automobile and airplane receivers. In such circumstances it is sometimes necessary to shield the entire ignition system, even including the spark plugs.

An expedient that has been successful in permitting reception in the presence of highly damped pulses, such as originate from ignition systems, motor brushes, power-line corona, and switches, is illustrated in Fig. 323. Here an auxiliary intermediate-frequency amplifier is connected with its grid in parallel with the grid of the final intermediate-frequency amplifier tube, and delivers its output to an auxiliary rectifier. The direct-current output of this auxiliary rectifier is then used to bias the third grid electrode of the final intermediate-frequency amplifier tube. With proper adjustment, a noise voltage appreciably greater than the signal being received will develop enough bias to make the intermediate-frequency tube inoperative, thus silencing the receiver for the duration of the noise.[1] This arrangement takes advantage of the fact that many noise voltages are of very short duration, with comparatively long quiet intervening intervals. With the arrangement shown in Fig. 323 the receiver output is simply turned off during these brief noise periods, and, although this causes some distortion, the reception is on the whole much improved.

The noise level in a radio receiver is usually increased by the presence of an unmodulated carrier, and, in fact, such a carrier wave often appears to be accompanied by a hiss which rises in intensity as the receiver is brought into resonance with the unmodulated carrier. This behavior results from heterodyne action between relatively weak noise voltages, such

[1] For further information see J. J. Lamb, A Noise Silencing IF Circuit for Superheterodyne Receivers, *QST*, vol. 20, p. 11, February, 1936.

as are normally present in the set, and the relatively strong unmodulated carrier wave. The loudness of the noise depends to a considerable extent on the amplitude of the carrier and is relatively small when the carrier is absent because of the decrease in detector efficiency which takes place when the removal of the relatively strong carrier causes the detector to change over from a power to a weak-signal rectifier.

Microphonic Action and Acoustic Feedback.—When the receiver chassis is in the vicinity of the loud-speaker, as is the case with the ordinary home radio receiver, there is always the possibility of sound vibrations affecting

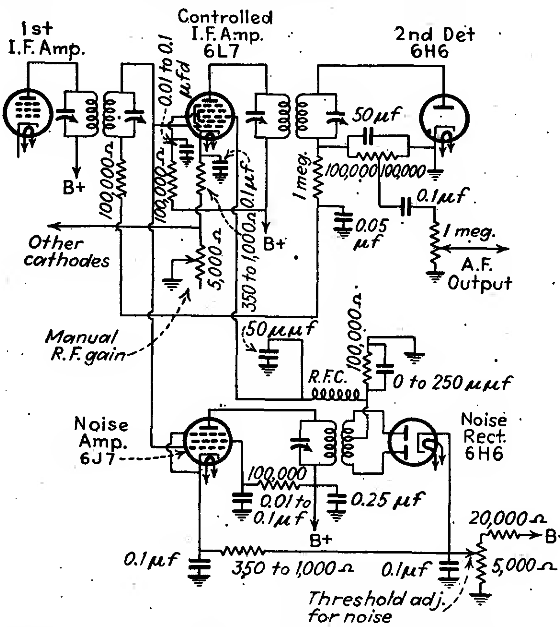


Fig. 323.—Noise-suppressing system in which a noise pulse stronger than the desired signal can be made to silence the receiver momentarily.

the radio receiver in such a manner as to produce acoustic feedback (*i.e.*, audio-frequency oscillations sustained by sound vibrations modulating an incoming carrier to produce still more sound output). Troubles of this type are particularly common with all-wave receivers delivering a large amount of audio power to the loud-speaker, and may arise from vibrations transmitted to the receiver chassis either through the air or through the cabinet material. The parts of the receiver most commonly giving trouble are the plates of the tuning condenser and the radio-frequency and oscillator coils. Acoustic feedback troubles can be kept under control by using rigid construction, by mounting the chassis on rubber supports, and by protecting the tuning coils and condenser as required, by rubber mounting. In console models the chassis can be

mounted on a shelf that acts as protection against sound waves. The receiver of Fig. 318 illustrates these usual expedients, and also obtains still further reduction in acoustic feedback by mounting the loud-speaker baffle, together with the loud-speaker, upon rubber to prevent vibrations from being transmitted directly from the loud-speaker to the cabinet.

Receivers that are to be subjected to intense mechanical vibration, such as airplane and automobile radios, must be rigidly constructed and are preferably mounted on rubber. It is also desirable to enclose such receivers completely to prevent sound waves from striking the coils, condensers, and tubes.

Special Problems Encountered in High-fidelity[1] *Receivers.*—High fidelity requires that modulation frequencies up to at least 10 kc be reproduced without excessive frequency distortion. This requires a good audio-frequency system and intermediate-frequency-amplifier and radio-frequency circuits that will respond to a band width approximately 20 kc wide. A receiver designed to meet these requirements is not entirely satisfactory for the reception of distant stations or reception in the presence of bad noise and interference. This is because the amount of noise that is received is directly proportional to the width of the frequency band that is received,[2] and because weak interfering signals on the adjacent channel will produce frequency components that extend into the response band of the receiver. This situation has led to the development of means for varying the width of the received band so that wide-band high-fidelity reception can be employed with strong local signals, while a narrow band is available for weak signals or when interference is bad. The simplest way of obtaining variable band reception is to vary the response band of the intermediate-frequency amplifier. One method of changing the band is discussed in Sec. 19, while another method is used in the receiver of Fig. 313. In the latter case, an intermediate-frequency coupling system having three tuned circuits is used, and the response band is controlled by a resistance in series with one of these circuits.[3]

[1] Further discussion is given by Harold A. Wheeler and J. Kelly Johnson, High Fidelity Receivers with Expanding Selectors, *Proc. I.R.E.*, vol. 23, p. 594, June, 1935; Stuart Ballantine, High Quality Radio Broadcast Transmission and Reception, *Proc. I.R.E.*, vol. 22, p. 564, May, 1934; vol. 23, p. 618, June, 1935.

[2] A wide response band not only increases the noise energy in the output but also causes the noise that is present to be more disturbing. Experiments show that, when the band width of an ordinary receiver is doubled to obtain a high-fidelity characteristic, the signal energy must be increased about ten times if the annoyance from noise is to be the same. See C. B. Aiken and G. C. Porter, Receiver Band Width and Background Noise, *Radio Eng.*, vol. 15, p. 7, May, 1935.

[3] It has been suggested that the band width might be automatically controlled by the strength of the received signal, thus giving automatic selectivity control (called A.S.C.). A discussion of means to carry out the necessary operation is given by H. F. Mayer, Automatic Selectivity Control, *Electronics*, vol. 9, p. 32, December, 1936.

In high-fidelity receivers that reproduce modulation frequencies up to 10 kc, trouble is commonly encountered from a 10-kc whistle produced by a weak adjacent-channel station heterodyning with the desired carrier frequency. This can be eliminated either by making the audio-frequency-system cut-off just below 10 kc or by the use of a 10-kc trap which removes a very narrow band of frequencies in the neighborhood 10,000 cycles, as in Fig. 313.

Amplitude distortion must be kept to an unusually low level in high-fidelity receivers if the full benefits of high fidelity are to be realized. This is because the amount of distortion that will produce a noticeable effect is less the greater the frequency band that is reproduced.[1] Furthermore, a high-fidelity system reproduces high-order harmonics (*i.e.*, those above about 4500 cycles), which are suppressed in an ordinary broadcast receiver and which have a particularly disagreeable effect.

The tuning indicator employed in a high-fidelity receiver must take into account the fact that the receiver responds almost equally well to carrier frequencies over a 20-kc range, whereas for proper operation the carrier must be adjusted to the exact center of the response band. This makes it essential that the tuning indicator operate from a separate sharply resonant circuit. Such an arrangement is shown in the high-fidelity receiver of Fig. 313. The importance of accurate tuning with high-fidelity receivers makes the use of automatic frequency control especially desirable with such receivers.

Some General Design Considerations.—The actual electrical design of commercial receivers depends to considerable extent upon such merchandising requirements as "the most receiver for a given selling price," or "the least expensive six-tube receiver," etc., and is influenced very considerably by the particular sales features that are being emphasized at the time. There are, however, certain general considerations governing the design. The critical factors are the audio-frequency power output, the sensitivity, and the carrier amplitude desired at the detector for normal operation. The audio power tubes and plate-supply voltage are selected to develop the required output. The minimum permissible audio-frequency amplification between detector and power tubes is the gain that will provide full excitation for the power tubes when the carrier voltage at the detector is completely modulated and has an amplitude equal to the normal detector level.[2] Starting with this carrier voltage, which will develop the full receiver output when completely modulated, it is a simple matter to determine the carrier amplitude modulated 30 per cent which

[1] See Frank Massa, Permissible Amplitude Distortion of Speech in an Audio Reproducing System, *Proc. I.R.E.*, vol. 21, p. 682, May, 1933.

[2] Actually most present-day receivers provide appreciably more audio gain than this minimum.

will develop the standard output of 50 mw. The total voltage step-up between antenna and second detector must then be such that, with a voltage corresponding to the required sensitivity acting in the antenna circuit, the required carrier level will be developed at the second detector. From this it is possible to estimate the number of stages of amplification required on the basis of gains to be expected according to Table XIV.

TABLE XIV.—TYPICAL GAINS IN COMMERCIAL RECEIVERS

Radio-frequency section:	
Antenna to grid of first tube[1]	3– 6
Radio-frequency amplifier, broadcast band	20– 40
Radio-frequency amplifier, short-wave bands	5– 25
Intermediate-frequency amplifier section:[2]	
Converter grid to first intermediate-frequency tube	40– 60
Interstage intermediate-frequency	50–180
Intermediate-frequency to diode	50–125

[1] In automobile receivers designed to operate with a particular antenna of specified constants, it is possible to make the antenna-to-grid step-up 20 to 50.

[2] These values are for air-cored coils for 460 kc. For air-cored coils at 260 kc the gains will be perhaps 30 per cent higher. With iron-dust cores the gains will be 20 to 60 per cent higher than with air cores.

The actual gains in any particular case depend upon the size and design of the coils and upon the coupling system involved. In the radio-frequency section the gain varies with the tuning, normally being greatest at the high-frequency end of each tuning range. In the broadcast band (550 to 1500 kc) the performance of both antenna and interstage parts can be improved by using a complex coupling arrangement with a high-inductance primary resonated at a frequency just below the broadcast band, as illustrated in Fig. 124. Such arrangements cannot be used in the short-wave bands, however, because the primaries would resonate with signals in the next lower band. Short-wave radio-frequency coils hence ordinarily employ a simple coupling system consisting of a low-inductance primary. The antenna-to-first-grid gain is necessarily low in ordinary broadcast receivers because these receivers must be designed so that the first tuned circuit will maintain its alignment when antennas of widely different constants are used, and to do this the antenna coupling must be small.

112. Receivers for Telegraph Signals.—Receivers for handling telegraph signals differ from broadcast receivers in that some means must be provided for interrupting the code characters at an audible rate so that reception can be obtained with a telephone receiver.[1] This is universally

[1] The only exceptions to this are when the code signals are produced by spark transmitters, or are in the form of interrupted continuous waves, or are to be recorded on a tape recorder. Spark and I.C.W. signals are interrupted at the transmitter and so can be received on an ordinary broadcast receiver. In the case of tape recording it is possible to operate the inking mechanism without producing an audible tone, although most of the commercial systems of tape recording involve an audible note which is rectified to operate the mechanism.

accomplished by the heterodyne principle, the procedure being to combine the incoming signal with a locally generated oscillation differing in frequency by about 1000 cycles. Rectification of the combination gives the code characters with a frequency of about 1000 cycles, as explained in Sec. 87.

One of the chief merits of the heterodyne method of reception is that the beat or difference frequency which is produced depends upon the frequency of the signal being received. This makes it possible to distinguish between signals which differ so little in frequency that ordinary tuned circuits cannot separate them. For example, if it is desired to receive from a station transmitting on 20,000,000 cycles while an interfering station is operating on a frequency of 20,000,200 cycles, it would be obviously impossible to eliminate the undesired signal by means of tuned circuits since the two carrier frequencies differ by only 0.001 per cent.

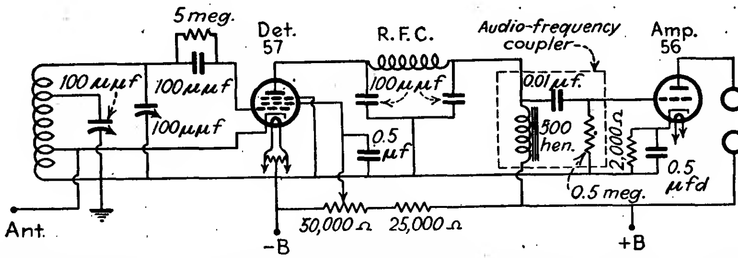


FIG. 324.—Simple code receiver consisting of an oscillator detector followed by one stage of audio-frequency amplification.

However, heterodyning with a local oscillation having a frequency of 20,001,000 cycles produces beat notes with the desired and undesired signals of 1000 and 800 cycles, respectively, and these frequencies can be readily separated either by ear or by a selective circuit because they differ by 25 per cent.

A simple code receiver is shown in Fig. 324, and consists of an oscillating detector of the type discussed in Sec. 89 followed by one stage of audio-frequency amplification. Such an arrangement possesses remarkable sensitivity because of the very large regenerative amplification obtained with a properly adjusted oscillatory detector. In some cases the receiver of Fig. 324 is modified by placing a radio-frequency tube ahead of the oscillating detector, using either a tuned or untuned coupling between the antenna and the grid of this tube. Such a tube is primarily for the purpose of making the local oscillator frequency independent of the receiving antenna and for preventing the radiation of local oscillator energy, but it also adds somewhat to the sensitivity.

Any ordinary superheterodyne receiver can be converted into a code receiver by the addition of a local oscillator for the purpose of heterodyning with the intermediate frequency. This fixed oscillator is detuned approxi-

mately 1000 cycles from the intermediate frequency and is loosely coupled into the circuits of the second detector, as illustrated in Fig. 325, thereby causing the latter to develop a beat note approximately 1000 cycles with the incoming carrier wave. If the superheterodyne receiver is provided with automatic volume control, this must be disconnected in order to prevent the local oscillator from affecting the receiver sensitivity, and a manual volume control must be substituted.

Compared with an oscillating detector receiver, the superheterodyne receiver has somewhat greater sensitivity, has less cross-talk trouble with strong local signals, and can be converted into an excellent receiver for telephone signals by simply turning off the local oscillator.

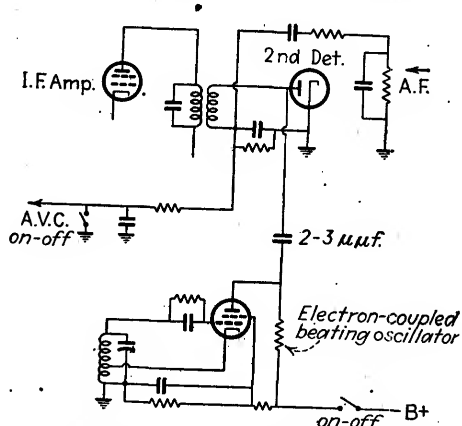


FIG. 325.—Circuit illustrating how a superheterodyne receiver may be converted into a code receiver by coupling a beating oscillator into the intermediate-frequency system.

Single-signal Receiver.—The single-signal receiver is a superheterodyne code receiver in which a crystal filter is used in the intermediate-frequency amplifier to obtain extremely high selectivity. The basic circuit is shown in Fig. 326 and makes use of a piezo-electric quartz crystal as a coupling element between two conventional intermediate-frequency tuned circuits. The condenser C_n is for neutralizing the electrostatic capacity of the crystal, and with perfect neutralization the crystal becomes equivalent to a series-resonant circuit that has a very high Q. By varying the neutralizing condenser C_n slightly it is possible to obtain a residual reactance in parallel with the crystal. This causes the crystal to go into parallel resonance at a frequency very close to the series-resonant frequency, and, because of the high impedance of a parallel circuit compared with a series circuit, the result is very little coupling at a frequency differing only slightly from that required for maximum coupling. The difference between this frequency of high attenuation and the frequency of maximum

response is controlled by the adjustment of C_n, and can be readily made as low as 1000 to 2000 cycles.

The width of the response band of the crystal filter with its associated circuits is largely controlled by the resistance which the crystal sees when looking back toward the circuit that excites it. This is because this resistance is effectively in series with the crystal, and so reduces the equivalent Q of the combination. The band width can accordingly be controlled by the condenser that varies the resonant frequency of the input circuit (see Fig. 326). When adjusted to resonance with the crystal, the effective circuit resistance is high and the maximum band width results, while, when the input circuit is appreciably detuned, the impedance that faces the crystal is low, and mainly reactive, so that the effective Q is very nearly the actual Q of the crystal itself. It is possible in this

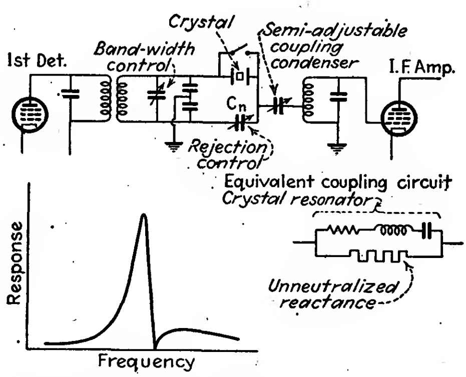


FIG. 326.—Crystal coupling arrangement used in single-signal receivers together with response curve illustrating how the image interference can be suppressed.

way to vary the band width from perhaps 100 cycles to 5000 cycles as the occasion requires.

Single-signal radio receivers greatly reduce interference. Thus signals differing in frequency from the desired signal by more than about 100 cycles are prevented from reaching the second detector by the great selectivity of the crystal filter, while in the ordinary code receiver all signals having a frequency within about 5000 cycles of the desired signal will cause audible beat notes. The single-signal receiver will also eliminate image effects produced at the final beating oscillator if the condenser C_n is adjusted so that the frequency of high attenuation coincides with the image frequency. Thus, if the intermediate frequency is 450 kc and the beating oscillation is 451 kc, signals producing an intermediate frequency of 452 kc are prevented from developing a 1000-cycle interfering beat note by adjusting the condenser C_n so that this image signal is suppressed before reaching the second detector.

Frequency Stability of Beating Oscillators.—In order to obtain satisfactory heterodyne reception it is necessary that the transmitted frequency and the frequency of the local oscillator, or oscillators, be constant, for otherwise the pitch of the audible beat note will vary and reception will be difficult. This is particularly true with single-signal receivers because here a small shift in the frequency of the first oscillator of the superheterodyne receiver will cause the intermediate frequency produced to shift clear out of the narrow response band of the crystal filter. The importance of frequency stability increases as the frequency of transmission increases because the same percentage of variation then represents a greater number of cycles. At extremely high frequencies, such as 100,000,000 cycles, it is difficult to obtain a steady beat note because of the effect of minute vibrations, etc., on the frequency.

113. Miscellaneous Types of Receivers. *Radio Telephone Receivers for Other Than Broadcast Purposes.*—Telephone receivers for other than broadcast purposes are in the main similar to broadcast receivers but with modifications of the tuning range and of the design details. Thus many airplane receivers are similar to automobile radios in the general method of construction, in the use of remote-control tuning, and in that a storage battery is the source of power. The chief differences are that the airplane receiver requires somewhat sturdier construction, covers only the airplane bands, and is arranged to operate a telephone receiver instead of a loud-speaker. In other cases airplane receivers are arranged so that only two frequencies can be received, one being the short-wave day frequency and the other the short-wave night frequency assigned to the air line involved. Superheterodyne receivers for such two-band operation commonly employ crystal oscillators for the frequency-changing process, with a different crystal for each frequency.

Receivers for long-wave telephone signals may be of the tuned radio-frequency type, such as illustrated in Fig. 314, but with perhaps two or three stages of radio-frequency amplification, or may be of the superheterodyne type having an intermediate frequency that either is quite low or is higher than the highest frequency to be received.

The ordinary superheterodyne receivers tend to give relatively poor image suppression at short waves unless an excessive number of tuned circuits is employed in the radio-frequency section. This difficulty can be overcome by employing a triple detection receiver as illustrated schematically in Fig. 327. The first intermediate frequency is high enough to provide adequate image suppression with one or two resonant circuits in the radio-frequency section, while the second intermediate frequency is commonly the same as in broadcast receivers and makes it possible to discriminate against signals differing only slightly from the desired frequency. The second oscillator of a triple-detection receiver

operates at a fixed frequency determined by the two intermediate frequencies, so that the only circuits which must be adjusted for the incoming signal are the radio-frequency and first oscillator sections.

Reception of Single Side-band Signals.[1]—Signals consisting of only a single side band can be received by using a local oscillator to supply the missing carrier frequency. This oscillator heterodynes with the frequency components of the single side-band signal, and by detection of the resulting combination the difference frequencies, which correspond to the original modulation frequencies, are reproduced.

The local oscillator which resupplies the carrier that was suppressed at the transmitter must be within 20 cycles of the correct frequency for reasonably intelligible speech, and even closer to the correct value if high quality is to be obtained. At frequencies below 1000 kc it is a simple matter to maintain the required tolerances by the use of well-designed and carefully supervised crystal oscillators at both transmitter and receiver. At higher frequencies it is necessary, however, to transmit to

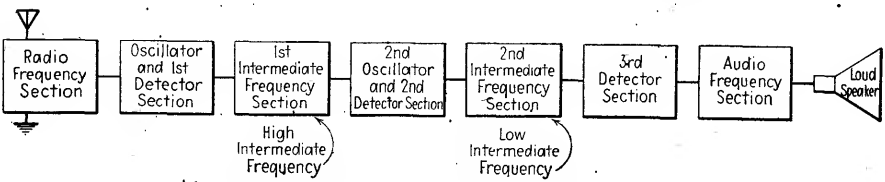


FIG. 327.—Schematic diagram of triple-detection receiver.

the receiver a pilot frequency which is most conveniently a small remnant of the actual carrier frequency of the transmitter. This pilot frequency may be used to synchronize the local oscillator automatically at the proper value, or may be separated from the side band, amplified, and used in place of the beating oscillator. The beating current obtained by the latter arrangement is termed a "reconditioned" carrier, and is entirely satisfactory provided the carrier amplifier has automatic volume control that maintains constant output irrespective of the fading of the residual carrier actually received. In either case the separation of the residual carrier from the single side band is done at the intermediate frequency, using a quartz-crystal filter having sufficient selectivity to suppress the side band while transmitting the residual carrier. In order to take care of frequency drifts either in the transmitted frequency or in the beating oscillation at the receiver, an automatic frequency-control arrangement is employed in the receiver to maintain the intermediate frequency at exactly the value required by the crystal filter.

[1] For further details concerning receivers for the reception of single side-band signals, see F. A. Polkinghorn and N. F. Schaack, A Single Side-band Short-wave System for Transatlantic Telephony, *Proc. I.R.E.*, vol. 23, p. 701, July, 1935.

Receivers for Ultra-high Frequency Signals.—In the wave-length range from ½ to 10 meters it is possible to use ordinary tuned radio-frequency and superheterodyne receivers. Acorn tubes are desirable in the high-frequency circuits for the shorter wave lengths in this range, and will produce a gain in radio-frequency amplification of 10 to 15 at wave lengths as short as 3 meters. When superheterodyne reception is employed, it is usually desirable to make the intermediate-frequency response band relatively wide in order to allow for some frequency instability in the transmitter and beating oscillators. Triple-detection superheterodyne receivers are favored when there are interfering stations to give image response.

Superregenerative detectors are often employed in receivers operating at wave lengths below 10 meters. The simplest arrangement of this type consists of a superregenerative detector followed by one or two stages of audio-frequency amplification. In more elaborate arrangements one or two tuned radio-frequency stages are placed between the antenna and the superregenerative detector in order to increase selectivity and prevent radiation from the receiver. Superregenerative receivers are remarkably sensitive, and have the advantage of producing nearly as much audio-frequency output with very weak signals as with strong signals, so that there is an inherent tendency to discriminate against ignition and other similar interference that is in the form of pulses of high intensity but short duration. The chief disadvantage of the superregenerative arrangement is that a strong hiss is present except when a carrier wave is being received. This hiss represents the inherent noise of the receiver caused by thermal agitation, etc.

Superheterodyne receivers for very short waves are sometimes provided with a superregenerative second detector because of the ability of such a detector to suppress ignition noise. For best results the intermediate frequency of such a superheterodyne receiver should be very high.

At wave lengths less than about ½ meter ordinary receivers fail to function properly, and no very satisfactory substitute has been found. The usual arrangement is some form of simple detector consisting either of a crystal rectifier, vacuum-tube rectifier, or electron-oscillator detector of the type discussed in Sec. 85. One or more stages of audio-frequency amplification can be added to increase the sensitivity.

Reception of Frequency-modulated Signals.[1]—Signals that are frequency (or phase) modulated can be received by discriminating against one side

[1] For further information on the reception of frequency-modulated waves, see J. G. Chaffee, The Detection of Frequency-modulated Waves, *Proc. I.R.E.*, vol. 23, p. 517, May, 1935. Also see E. H. Armstrong, A Method of Reducing Disturbances in Radio Signalling by a System of Frequency Modulation, *Proc. I.R.E.*, vol. 24, p. 689, May, 1936.

band so that the carrier and remaining side band heterodyne together to give beat notes corresponding to the original modulation.[1] Any ordinary receiver that has high selectivity can accordingly be used to receive frequency-modulated signals by detuning the receiver so that the carrier is on the side of the response curve as illustrated in Fig. 328. The most suitable arrangement is a superheterodyne receiver with the intermediate-frequency amplifier designed to give the desired selectivity characteristics.

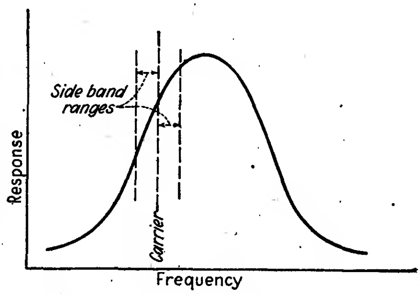


FIG. 328.—Response curve of receiver showing how one side band of a frequency-modulated wave can be discriminated against by detuning the receiver slightly.

If the output of the receiving system is to be proportional to the amplitude of the modulating voltage at the transmitter, it is necessary that the amplitude of the side-band components be inversely proportional to the frequency of the modulating voltage. This is because the discrimination that the selectivity curve of the receiver is able to exert between the upper and lower side bands is proportional to the modulation frequency, causing the receiver to discriminate against the lower modulating frequencies. This introduces no difficulty with frequency modulation, since from Sec. 81 it is seen that the modulation index for a frequency-modulated wave is inversely proportional to the modulation frequency. However, with phase modulation the amplitude of the modulating voltage must be made inversely proportional to modulating frequency by means of a network if excessive frequency distortion is to be eliminated at the receiver. The amplitude distortion produced in the reception of frequency-modulated signals tends to be comparatively high unless the modulation index of the transmitter is low and the portion of the selectivity curve made use of at the receiver is substantially linear.

114. Receiving Systems for Minimizing Fading.—Fading commonly occurs at broadcast and higher frequencies, and is the result of interference between waves that have traveled to the receiver along different routes, as explained in Chap. XIV. When fading is present, the signal intensity varies widely in a random manner, and with radio-telephone

[1] The reception of frequency-modulated signals by detuning the receiver is sometimes explained on the basis that, as the frequency of the signal is varied, the response of the radio-frequency circuits varies likewise, thus converting the frequency-modulated wave into an amplitude-modulated wave. This view is only partially correct, however, since the frequency-modulated wave, instead of being a wave of varying frequency, is really a wave containing carrier and side-band components as explained in Sec. 81.

signals there is usually marked distortion during the weak part of the fading cycle.

Automatic volume control is an obvious means of minimizing the effects of fading, and is used in nearly all broadcast receivers. This is not a complete solution of the problem, however, because the signals sometimes fade out so completely as to be lost in the background noise, in which event no amount of amplification will be successful in retrieving them. Also, as the signals fade in and out, the sensitivity of the receiver is varied, and there is a continual variation in the background noise. Furthermore, automatic volume control does not prevent the quality distortion usually accompanying fading, and cannot be employed with code receivers because of the violent variations in receiver gain that would result when the transmitter keying is momentarily interrupted.

Diversity Systems for Code Reception.—Experiments have shown that frequencies differing by a few hundred cycles do not fade simultaneously, and this makes it possible to minimize fading troubles by transmitting the same message on different frequencies, *i.e.*, by frequency diversity. Many short-wave code transmitters accordingly modulate the transmitter output at some audible frequency, say 500 cycles, and then key this modulated wave. The audio modulation introduces side-band frequencies and in effect causes the telegraph message to be transmitted simultaneously on a number of frequencies, which is of considerable help in reducing the number of drop-outs.

A still better anti-fading system can be obtained by taking advantage of the experimentally observed fact that the signals which are induced in antennas spaced 10 wave lengths or more apart do not fade simultaneously. When three or more such antennas are employed, it is extremely unlikely that the signals will fade out completely on all of them at the same time, so that, if the outputs of the different antennas are combined, there will nearly always be some signal present.

An example of a commercial installation making use of a spaced-antenna receiving system for minimizing fading of code signals is shown schematically in Fig. 329.[1] This makes use of three short-wave antennas spaced 10 wave lengths apart and each provided with a separate code receiver. The audio outputs are rectified as shown and combined by passing through a common resistance. The voltage drop across this resistance is amplified by a one-stage direct-current amplifier, the output of which is passed through a resistance that supplies the grid-bias voltage for a push-pull amplifier. The input of this amplifier is excited from an audio-frequency "tone" oscillator. When no signals are being received,

[1] For a detailed description of this particular installation, see H. H. Beverage and H. O. Peterson, Diversity Receiving Systems of RCA Communications, Inc., for Radio Telegraphy, *Proc. I.R.E.*, vol. 19, p. 531, April, 1931.

the current through the resistance common to the rectified outputs of the three receivers is negligible, causing the plate current of the direct-current amplifier tube to be large. This places a rather high negative bias on the push-pull amplifiers, blocking them. If, however, a signal is received from at least one of the antenna systems, current flows through the com-

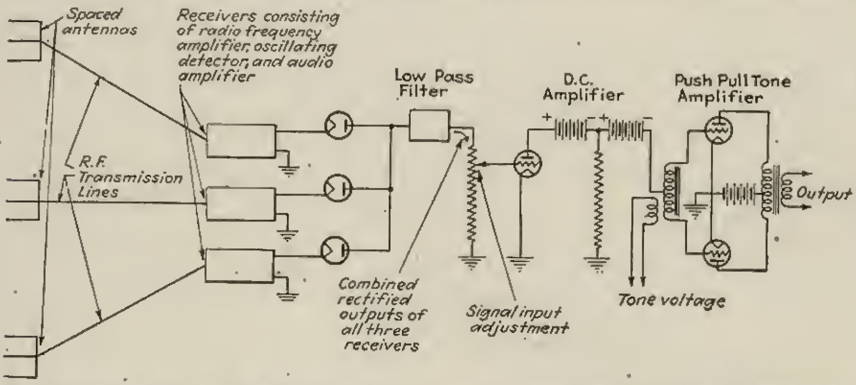


FIG. 329.—Diagram showing operation of typical diversity receiving system for telegraph signals.

mon output resistance of the three receivers, making the grid of the direct-current amplifier more negative, reducing the amplifier plate current, and lowering the grid bias of the push-pull amplifier. This enables the push-pull amplifier to pass the audio frequency from the tone oscillator. The adjustment on the grid bias of the push-pull amplifier is normally such that, when no signal is being received, the noise level will just fail to allow the push-pull amplifier to pass tone current.

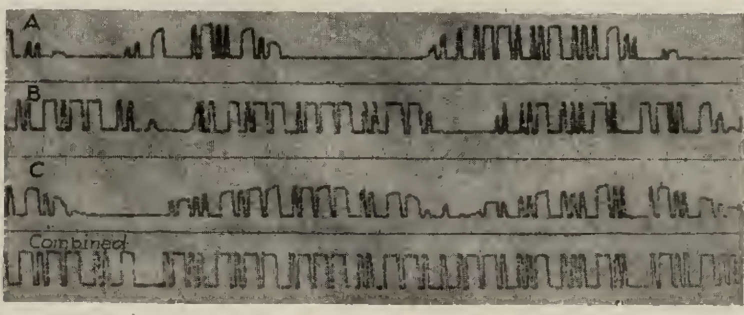


FIG. 330.—Typical records showing output of each receiver of a diversity receiving system and of the combined output. Note the entire absence of fading in the latter.

A diversity arrangement as described will eliminate virtually all drop-outs resulting from fading, since it only rarely happens that the signals will fade out completely on all three receiving systems at the same instant. This is well illustrated by the records of Fig. 330, which compare the combined output of the diversity receiving system with the

individual outputs of each antenna. The arrangement by which the output of the receiving system keys a local oscillator gives a final output free of noise and having an amplitude and frequency independent of the adjustments of the radio receiver.

Diversity Receiving Systems for the Reception of Radio-telephone Signals.—Most fading troubles in the reception of radio-telephone signals can be eliminated by using a spaced-antenna system and by taking advantage of the observed fact that, when the received carrier is strong, the quality distortion is normally a minimum.[1] The outputs of the spaced antennas are accordingly combined in such a manner that the antenna with the strongest carrier always contributes a disproportionately large share of the total output. One means of doing this is shown schematically in Fig. 331.[2] Here the outputs of three receivers having square-law detectors and excited from three spaced antennas are combined directly as shown. With this arrangement the antenna receiving the strongest carrier contributes nearly all the output because the square-law detector favors the stronger signal.[3] The receivers, which at the moment are contributing little to the output, are prevented from contributing to the noise level by the use of an automatic-volume-control voltage obtained from the common rectified output as shown and operating simultaneously on all receivers. When the output of any one receiver is strong, the control reduces the amplification of all three receivers and hence keeps down the noise in the common output.

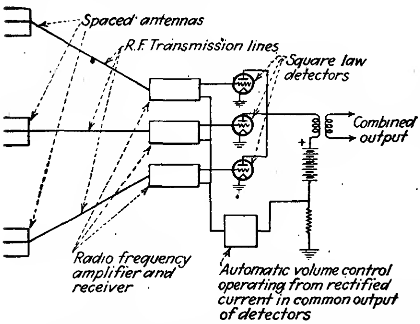


FIG. 331.—Schematic diagram of diversity telephone receiving systems. The outputs of the square-law detectors are combined directly, and an automatic volume control operates on all receivers.

[1] This is because the worst quality distortion is that caused by the carrier fading out while leaving relatively strong side bands.

[2] This is the diversity receiving system used by the RCA Communications, Inc., at their Riverhead, Long Island, receiving station. For a more detailed description see H. O. Peterson, H. H. Beverage, and J. B. Moore, Diversity Telephone Receiving System of RCA Communications, Inc., *Proc. I.R.E.*, vol. 19, p. 562, April, 1931.

[3] A still better arrangement would be one involving linear detectors, each followed by one stage of audio-frequency amplification. The outputs of the audio amplifiers would be combined, but a quieting system would cut off the output of all the audio amplifiers except the one associated with the detector receiving the strongest carrier. The control for the quieting systems could be obtained by a differentially operated system of tubes. Such an arrangement would avoid the distortion of the square-law detectors.

Directivity Steering.—The different paths followed by the signals when fading is present are found experimentally to arrive at the receiver at different vertical angles, and furthermore the signals following a particular path are relatively stable. This has led to the suggestion that fading could be minimized by employing an antenna system in which the vertical directivity was always adjusted to discriminate in favor of the waves arriving along the lowest vertical angle, whatever this happened to be. Experiments with an antenna having easily adjustable vertical directivity indicate that it is possible to obtain marked reduction in fading by the use of such directivity steering, and demonstrate the basic soundness of the sharp angular method of discrimination.[1]

Problems

1. Discuss the effects that each section of Fig. 309 can be expected to have on the fidelity of a broadcast receiver.

2. Discuss the factors that can cause the selectivity and fidelity characteristics of a superheterodyne receiver to vary with the frequency being received.

3. Make an estimate of the sensitivity of the receiver of Fig. 314, using reasonable assumptions as to stage gains and detector efficiencies.

4. Make a separate detail drawing of the second detector and volume-control system of the receiver of Fig. 310 and explain the function of each circuit component.

5. Explain the method used in the receiver of Fig. 310 to bias the grids of the audio and output tubes.

6. Make an estimate of the total d-c voltage and current that must be developed across the filter input condenser of the receiver of Fig. 310, assuming normal tube-operating conditions and using tube data from a tube manual.

7. Assuming that the speaker field of the receiver of Fig. 310 has an incremental inductance of 8 henries, calculate: (*a*) the ratio of alternating-current voltage in the plate supply of the output tube to the alternating-current ripple voltage appearing across the input condenser of the rectifier-filter system, (*b*) the ratio of alternating-current voltage appearing in the plate supply of the 6F5 audio tube to the alternating-current ripple voltage across the input condenser of the rectifier-filter system.

8. Make a rough estimate of the sensitivity of the receiver of Fig. 310 by calculating stage gains where possible and in other cases assuming reasonable gains.

9. Draw a detail diagram of the intermediate-frequency amplifier of the receiver of Fig. 312, showing the high-frequency filter and by-pass systems in plate, screen, and cathode circuits. Discuss the factors that determine the size of each condenser and resistance involved and calculate the effectiveness of the by-passing (or filtering) at the intermediate frequency of 450 kc.

10. Draw a detail diagram of the converter circuit used in the receiver of Fig. 312, and indicate the function of each circuit element.

11. Explain the operation of the bass-compensating system incorporated in the volume control of the receiver of Fig. 313.

12. In the receiver of Fig. 313 separate diodes are used for automatic volume control and for detection. Explain the operation of the circuit arrangement shown, paying particular attention to the means used to obtain delay for the automatic-

[1] See E. Bruce and A. C. Beck, Experiments with Directivity Steering for Fading Reduction, *Proc. I.R.E.*, vol. 23, p. 357, April, 1935.

volume-control system, while allowing normal detector action for the signal diode, and to the biasing system that enables the 6B7 tube to function as a normal amplifier.

13. Make a detail drawing of the quieting system used in Fig. 313 and explain the action by reference to this diagram.

14. Make a detail drawing of the tuning indicator in Fig. 313 and explain the operation. Include in the explanation the means by which the indicator can be made to tell when the receiver is adjusted to the exact center of the response band even though the receiver gives the same output over an appreciable frequency range.

15. A radio-frequency tuning coil is to cover the frequency range 530 to 1650 kc with a maximum capacity of 380 $\mu\mu$f:

a. Calculate the coil inductance that is required and the total minimum circuit capacity.

b. If the coil in *a* is to be used in a superheterodyne receiver having an intermediate frequency of 450 kc, calculate oscillator inductance and oscillator series and shunt trimmer condensers (assuming shunt trimmer is in parallel with the tuning condenser) which will give the desired intermediate frequency when the radio frequency coil is tuned to 600, 1000, and 1400 kc.

c. Calculate a curve similar to that of Fig. 316 showing accuracy of tracking for the design in (*b*).

16. A receiver covering the frequency range 530 to 1650 kc has an intermediate frequency of 460 kc. List the conditions and frequencies at which whistles or spurious responses are most likely to occur.

17. Design a superheterodyne receiver for the broadcast band (550 to 1500 kc) which will have a sensitivity of the order of 30 μv. This design is to be worked out in detail with constants given for all circuit elements, or is to be given only in a general form with preliminary calculations, tube line-up, and simplified schematic circuit, according to the assignment.

18. *a.* In a superheterodyne code receiver it is permissible to make the receiver selectivity much greater than when radio-telephone signals are to be received. Explain why this is so.

b. Explain why the use of the highest possible selectivity in a code receiver makes it possible to receive weaker signals than when less selectivity is used.

19. Work out the circuit of a code receiver consisting of one stage of tuned radio-frequency amplification, an oscillating detector, and a two-stage audio-frequency amplifier.

20. *a.* In the heterodyne reception of code signals explain why stability of the transmitted frequency is just as important as stability of the beating oscillator of the receiver.

b. When a superheterodyne receiver is used in the reception of code signals, what advantages (if any) are to be gained by the use of automatic-frequency control for (1) the first oscillator of the receiver, and (2) for the second oscillator?

21. When an ordinary receiver is used to receive a frequency-modulated signal by tuning so that the carrier is on the side of the resonance curve, it is found that there are two carrier frequencies for which the reception of frequency-modulated waves is equally efficient, and there is also a third carrier frequency for which amplitude-modulated waves are received. Explain how these various responses occur, and suggest a means whereby the receiver can be modified to eliminate or greatly reduce the reception of all signals except one frequency-modulated carrier.

CHAPTER XIV

PROPAGATION OF WAVES

115. Ground-wave Propagation.—The energy radiated from a transmitting antenna located at the earth's surface can be conveniently divided into a ground wave, which travels along the surface of the earth, and a sky wave, which is propagated in the atmosphere above the earth. The ground wave starts from the transmitter with a strength determined by the field radiated along the horizontal, but as it travels away from the transmitting antenna it becomes weaker as a result of spreading and because of energy absorbed from the wave by the earth. These earth losses result from the fact that the wave induces charges in the ground as illustrated in Fig. 1, and as these charges travel along with the wave they represent a current that dissipates energy. The portion of the wave in contact with the earth is therefore being continuously and relatively rapidly wiped out, only to be replenished, at least in part, by a diffraction of energy downward from the portions of the wave immediately above the earth. As a result of this situation, the strength of the ground wave is attenuated with distance at a rate determined by the distance, the frequency, the earth conductivity, and dielectric constant of the earth, in a complicated manner.

Sommerfeld Analysis of Ground-wave Propagation for Resistive Earth. The solution of the relations involved in ground-wave propagation was first given by Sommerfeld in a celebrated paper.[1] According to this, the strength of the ground wave produced by an antenna located at the surface of a flat earth is given by an equation of the form

$$\text{Ground-wave field strength} = \epsilon = \frac{k}{d}A \tag{206}$$

where

k = a constant determined by the strength of the field radiated along the horizontal by the transmitting antenna

d = distance from the transmitter

A = a factor that takes into account the ground loss.

[1] A. Sommerfeld, The Propagation of Waves in Wireless Telegraphy, *Ann Phys.*, vol. 28, no. 4, p. 665, 1909; vol. 81, p. 1135, 1926.

An excellent discussion of the Sommerfeld analysis, together with modifications that permit its application to practical problems, is given by K. A. Norton, The Propagation of Radio Waves over the Surface of the Earth and in the Upper Atmosphere, *Proc. I.R.E.*, vol. 24, p. 1367, October, 1936.

It will be noted that $\epsilon = k/d$ represents the strength of the field that would be obtained at the surface of the ground on the assumption of a perfect earth (*i.e.*, infinite conductivity). In the important special case where the antenna is vertical and is short compared with a wave length (*i.e.*, radiated field proportional to the cosine of the angle of elevation), $k = 300\sqrt{P}$ (where P is the radiated power in kilowatts), ϵ is in millivolts per meter, and d is in kilometers, or $k = 186.4\sqrt{P}$ when d is in miles.

The Sommerfeld reduction factor A depends in a relatively complicated way upon the frequency, dielectric constant and conductivity of the earth, and the actual distance. By making modifying assumptions that do not introduce appreciable error under the conditions existing in practical radio communication, it is possible, however, to express the reduction factor in terms of two parameters, the numerical distance p, and the phase constant b defined as follows:[1]

$$\tan b = (\varepsilon + 1)\left(\frac{f_{kc}}{1.8 \times 10^{18}\sigma}\right) \tag{207a}$$

$$\left.\begin{aligned} \text{Numerical distance } p &= \pi\left(\frac{f_{kc}}{1.8 \times 10^{18}\sigma}\right)\frac{d}{\lambda}\cos b \\ &= \frac{\pi}{\varepsilon + 1}\left(\frac{d}{\lambda}\right)\sin b \end{aligned}\right\} \tag{207b}$$

where

ε = dielectric constant of earth (assuming air is unity)
σ = conductivity of earth in electromagnetic units
f = frequency in kilocycles
d/λ = distance from transmitting antenna in wave lengths.

The relationship between p and b and the reduction factor A is quite involved, but is given graphically in Fig. 332, and can also be expressed with fair approximation by the empirical relation

$$\left.\begin{array}{l}\text{Sommerfeld reduc-}\\ \text{tion factor}\end{array}\right\} = A = \frac{2 + 0.3p}{2 + p + 0.6p^2} - \sqrt{\frac{p}{2}}\,e^{\frac{-5p}{8}}\sin b \tag{208a}$$

Discussion of the Sommerfeld Analysis.—The phase constant b depends upon the power factor of the soil when considered as a conductance shunted by a dielectric. At low frequencies, where the impedance offered by the earth to the flow of current is primarily resistive, the constant b is small and the numerical distance p then depends only upon the conductivity, frequency, and actual distance, and is independent of the dielectric constant of the earth. For this condition

[1] This follows the treatment of Norton, *loc. cit.*

$$\left.\begin{array}{l}\text{Numerical distance with}\\ \text{resistive earth}\end{array}\right\} = \pi\left(\frac{f_{kc}}{1.8 \times 10^{18}\sigma}\right)\frac{d}{\lambda} \tag{208b}$$

On the other hand, at high frequencies where the impedance offered by the earth to the flow of current is primarily capacitive, the phase constant b approaches 90°. The numerical distance then depends only upon

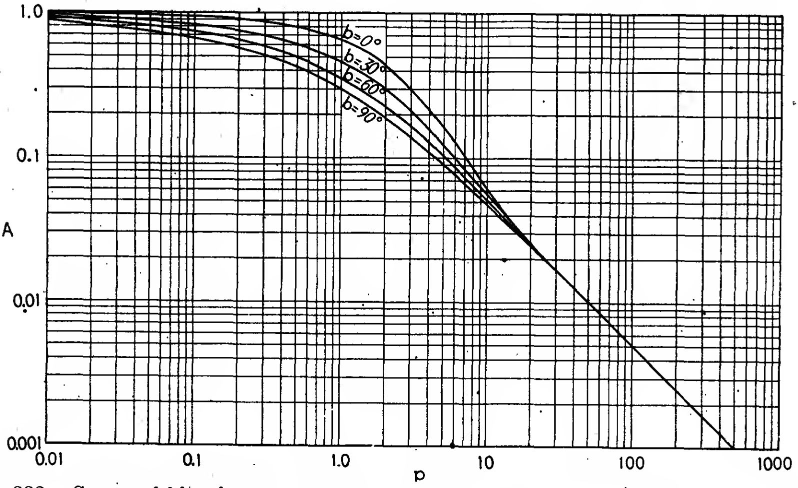


FIG. 332.—Sommerfeld reduction factor A as a function of the numerical distance p for various values of phase constant b.

the dielectric constant and the actual distance measured in wave lengths, and is independent of earth conductivity. For this case

$$\left.\begin{array}{l}\text{Numerical distance with}\\ \text{capacitive earth}\end{array}\right\} = \frac{\pi}{\varepsilon + 1}\frac{d}{\lambda} \tag{208c}$$

The frequency at which the soil changes over from mainly conductive to mainly capacitive impedance depends upon the soil constants, but with ordinary earth this begins to take place at the high-frequency end of the broadcast band (1500 kc), and the transition is ordinarily fairly complete at the frequencies used for long-distance short-wave communication.

The Sommerfeld attenuation factor A produced by the earth depends primarily upon the numerical distance p. The attenuation is small for numerical distances up to unity, but thereafter increases rapidly. For numerical distances up to approximately 10 the phase constant has some effect upon the attenuation, as shown in Fig. 332, but for greater numerical distances the attenuation factor approaches the value $1/2p$ and becomes independent of the phase constant. For such large distances the reduction factor A is inversely proportional to distance, and the

strength of the ground wave is then inversely proportional to the *square of the distance.*

At broadcast and lower frequencies where the earth is primarily resistive, the attenuation that the ground wave suffers at a given distance is particularly sensitive to the frequency and the conductivity. Examination of Eq. (207*b*) shows that, when b is small, the numerical distance corresponding to a given physical distance is inversely proportional to the earth conductivity and directly proportional to the square of the frequency. High ground conductivity and low frequency are then very important in obtaining strong ground-wave signals at considerable distances. At high frequencies the earth is primarily capacitive, and the numerical distance corresponding to a given physical distance is proportional to frequency and approximately inversely proportional to the dielectric constant of the earth, so that high dielectric constant and low frequency reduce the ground-wave attenuation.

The conductivity and dielectric constant effective in Eq. (207) are the average values for the earth extending down to depth of a few to over a hundred feet, according to the frequency. The dielectric constant and conductivity can be most satisfactorily derived from data on ground-wave propagation taken over the desired path at some convenient frequency. Knowing the frequency, distance, and rate of attenuation, it is possible to work backward and find the earth constants required to account for the results. The values so obtained can then be used to determine the ground-wave propagation of waves of other frequencies over the same general path.[1] In the case of homogeneous soil it is possible to employ sampling methods with fair success, although if any accuracy is to be realized this requires the averaging of a large number of samples.

The soil conductivity varies greatly under different conditions, ranging from about 3×10^{-13} electromagnetic units for moist loam to values of the order of 10^{-14} for dry sandy and rocky soil. The conductivity of salt water depends on the temperature and the saline content, and averages about 5×10^{-11} electromagnetic units. The dielectric constant of earth commonly ranges from 5 to 30, with the low values tending to go with dry soils of poor conductivity and the larger values with moist conducting earth. The dielectric constant of salt water is approximately 80.

The Sommerfeld analysis neglects the curvature of the earth and so gives calculated fields that are too high when the distance is great. The analysis of this case is quite complicated and of more theoretical

[1] Examples of such determinations are given by S. S. Kirby and K. A. Norton, Field Intensity Measurements at Frequencies from 285 to 5400 Kilocycles per Second, *Proc. I.R.E.*, vol. 20, p. 841, May, 1932.

than practical interest, since other factors such as fading and sky waves ordinarily become the dominant factors before the additional attenuation due to earth curvature becomes large.[1]

Curves showing ground-wave attenuation as a function of distance for different wave lengths and earth conductivities, and including the effect of earth curvature, are given in Fig. 333 for the case of a transmitting antenna radiating 1 kw of power and having a directional characteristic such that the strength of the radiated field is proportional to the cosine

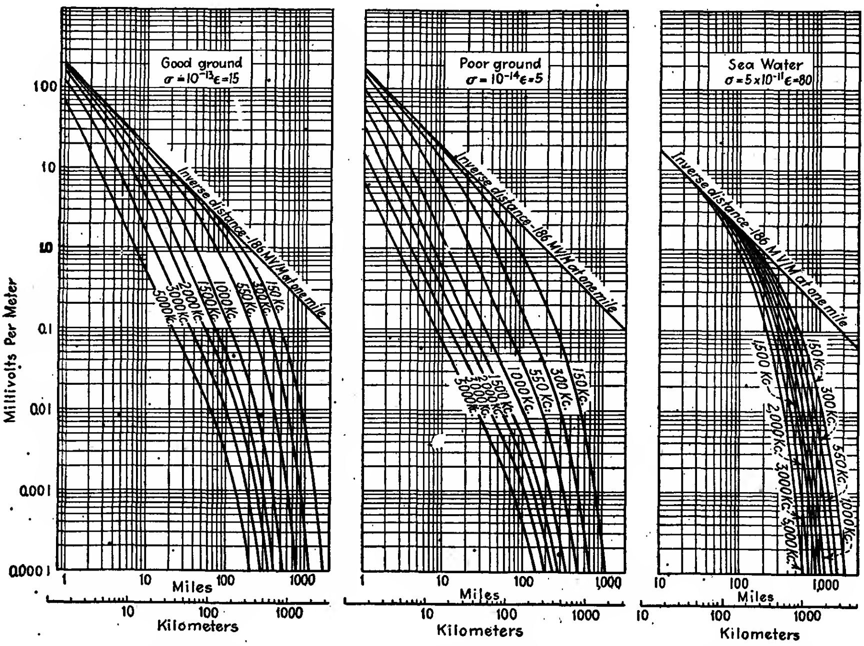


FIG. 333.—Strength of ground wave as a function of distance, frequency, and soil conductivity for 1 kw radiated power from an antenna producing a radiated field proportional to the cosine of the vertical angle. (For more complete curves see *Proc. I.R.E.*, vol. 21, p. 1419, October, 1933.)

of the angle of elevation above the horizontal.[2] For other powers the field strength will be directly proportional to the square root of the power, while with other directivities the field strength must be multiplied by a factor giving the relative horizontal field strength for 1 kw

[1] For further information on the effect of earth curvature, see Norton, *loc. cit.*; T. L. Eckersley, Direct Ray Transmission, *Proc. I.R.E.*, vol. 20, p. 1555, October, 1932; Charles R. Burrows, Radio Propagation over a Spherical Earth, *Proc. I.R.E.*, vol. 23, p. 470, May, 1935.

[2] These are taken from Norton, *loc. cit.* For additional curves see Report of Committee on Radio Propagation Data, *Proc. I.R.E.*, vol. 21, p. 1419, October, 1933.

radiated in the actual case compared with the field strength for the cosine directivity. Thus a 50-kw broadcast transmitter employing an antenna that increases the field strength along the horizontal to 1.41 times the field strength obtained with the assumed cosine law would make the radiated field $1.41\sqrt{50} = 10$ times the value obtained from Fig. 333. For purposes of comparison, inverse distance curves corresponding to zero earth losses are also shown. Examination of the curves of Fig. 333 shows the very great effect of frequency and earth conductivity upon the ground-wave field strength, and brings out clearly how low frequency and high ground conductivity improve the coverage of ground wave signals.

Ground-wave Propagation over Composite Paths.—When the earth conductivity differs appreciably at different places along the path of the ground wave, the Sommerfeld analysis cannot be expected to yield accurate results. Under such conditions it is necessary either to use an average earth conductivity or to piece together attenuation curves corresponding to the various soil conductivities. Neither method can be expected to give entirely satisfactory results, although the latter has the most rational basis.[1]

An important case of a composite path occurs when the transmitting station is located near the sea coast and is intended for transmission out to sea. Sea water has very low ground-wave absorption as compared with land, and it is accordingly very important that the transmitting station be located as close as possible to the water's edge. Experimental investigations have shown that, when a short-wave transmitter is located a mile inland from the shore, the additional attenuation of the ground wave is of the order of 10 db.[2]

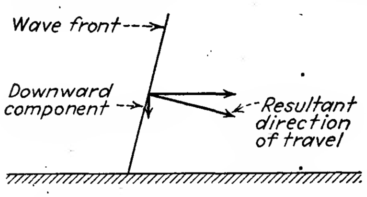


Fig. 334.—Wave front of ground wave showing the forward tilt that is a result of the horizontally traveling energy combining with the downward flow that replenishes the earth losses.

Character of Ground Wave.—The ground wave is always vertically polarized, and has a wave front that possesses a slight forward tilt. The vertical polarization exists because the conducting earth short circuits any horizontally polarized component of the wave, and hence wipes out the electrostatic flux of a horizontally polarized ground wave. The forward tilt is the result of the fact that energy is being continuously diffracted toward the earth to replenish the earth losses, as has already been explained, and this downward flow of energy, combined with

[1] A procedure for joining attenuation curves is given by P. P. Eckersley, The Calculation of the Service Area of Broadcast Stations, *Proc. I.R.E.*, vol. 18, p. 1160, July, 1930.

[2] See R. A. Heising, Effect of Shore Station Location upon Signals, *Proc. I.R.E.*, vol. 20, p. 77, January, 1932.

the horizontal travel along the surface of the earth, gives a resultant tilt as illustrated in Fig. 334.[1]

116. The Sky Wave.—The energy radiated in directions other than along the ground surface travels through space until reaching the ionized region in the upper atmosphere, where, if conditions are favorable, the path of the wave will be bent earthward. Such a sky wave may return to earth at very great distances from the transmitter, and is the means of obtaining long-distance radio communication. The sky wave in traveling through the space between the earth's surface and the ionized region suffers little attenuation other than that caused by spreading, so that the field strength in this portion of the path is inversely proportional to distance. In the ionized region there are additional losses which under some conditions will be sufficient to absorb the sky wave completely and which under others will introduce relatively little attenuation beyond that corresponding to the inverse-distance law.

117. Nature of the Ionosphere.[2]—Ultra-violet light or some other form of ionizing radiation originating with the sun, and apparently traveling with the velocity of light, ionizes the outer portion of the earth's atmosphere. This ionized region or *ionosphere*, also termed the Kennelly-Heaviside layer after the two scientists who independently and almost simultaneously suggested its existence, consists of free electrons, positive ions, and negative ions, in a rarefied gas. The effect

[1] For further information on the quantitative relations involving tilt angle see G. W. O. Howe, "The Tilt of Radio Waves and Their Penetration into the Earth," *Wireless Eng. and Exp. Wireless*, vol. 10, p. 587, November, 1933.

This paper also gives formulas for calculating the penetration of the radio-frequency currents into the earth.

[2] The literature on the ionosphere is very extensive, and this and the following several sections can do no more than summarize some of the principal points. The reader desiring to study this subject further is referred to the following selected bibliography which covers in a comprehensive way the knowledge concerning the ionized layer and includes reference to the remaining literature:

P. O. Pedersen, The Propagation of Radio Waves, *G.E.C. Gad*, Copenhagen, Denmark (in English); S. S. Kirby, L. V. Berkner, and D. M. Stuart, Studies of the Ionosphere and Their Application to Radio Transmission, *Proc. I.R.E.*, vol. 22, p. 481, April, 1934; E. O. Hulburt, The Ionosphere, Skip Distances of Radio Waves, and the Propagation of Micro Waves, *Proc. I.R.E.*, vol. 23, p. 1492, December, 1935; J. P. Schafer and W. M. Goodall, Diurnal and Seasonal Variations in the Ionosphere during the Years 1933 and 1934, *Proc. I.R.E.*, vol. 23, p. 670, June, 1935; Theodore R. Gilliland, Multi-frequency Ionosphere Recording and Its Significance, *Proc. I.R.E.*, vol. 23, p. 1076, September, 1935; S. S. Kirby and E. B. Judson, Recent Studies of the Ionosphere, *Proc. I.R.E.*, vol. 23, p. 733, July, 1935; L. V. Berkner and H. W. Wells, Report of Ionosphere Investigations at the Huancyo Magnetic Observatory (Peru) during 1933, *Proc. I.R.E.*, vol. 22, p. 1102, September, 1934; E. O. Hulburt, Wireless Telegraphy and the Ionization in the Upper Atmosphere, *Proc. I.R.E.*, vol. 18, p. 1231, July, 1930.

that the ionosphere has on radio waves is a result of the free electrons, and is determined by the distribution of electron density in the upper atmosphere.[1] The electron density is negligible in the earth's lower atmosphere, but reaches a value sufficient to influence radio waves at a height of about 100 km, and then maintains a relatively high value up to 300 to 500 km, after which the density presumably tapers off. The maximum electron density and the exact way in which the density varies with height depend upon the time of day and the season and also vary from year to year.

During the day there are three distinct maxima of electron density as illustrated in Fig. 335*a*. These are termed the E, F_1, and F_2 layers, as indicated, and have progressively higher maximum density. These regions of maximum electron density are not strictly layers as there is ionization between them as shown in Fig. 335, but the term layer is generally employed to designate the regions near the maxima. The height

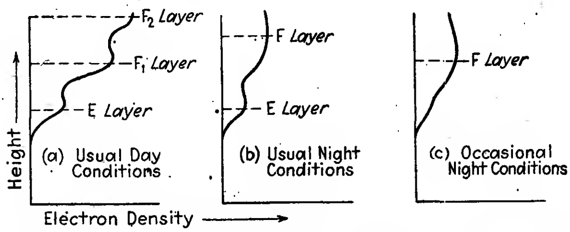


FIG. 335.—Schematic diagram illustrating the variation of electron density with height above earth under typical conditions.

of the maximum density of the E layer is substantially constant at about 100 km, with little diurnal or seasonal variation. For the F_1 layer, the height of maximum electron density is approximately 200 km, and this likewise has very little diurnal or seasonal variation. In contrast with this, the height of the F_2 layer has considerable diurnal and seasonal change and also depends upon the latitude, with typical heights lying within the range 250 to 350 km.

At night the F_1 layer tends to fade out while the F_2 layer descends, with the result that these two layers coalesce to form a single night-time F layer, as shown in Fig. 335*b*. This layer has a height of approximately 250 km and has little diurnal and seasonal variation. The E layer also fades at night, and in some cases may actually disappear, as indicated in Fig. 335*c*.

[1] In the lower, or E layer, there is the possibility that the ions are sufficiently numerous to have some influence. Under such conditions the term *electron density* represents the electron density that would have the same effect as the actual ionized medium.

This behavior of the ionosphere layers is illustrated in Fig. 336, which brings out the coalescing of the F_1 and F_2 layers at night and also shows the extent of the variability of the F_2 layer with season.

Electron Density.—Little is known about the distribution of the electrons in the ionosphere other than the height and maximum electron density of each layer. Other details can only be inferred on the basis of theoretical hypotheses.

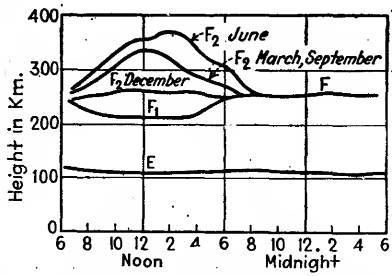


FIG. 336.—Average height of the ionized regions at Washington, D. C. during 1933–1934.

Typical behavior of the electron density of the different layers is illustrated by the curves of Fig. 337. The maximum density of the E and F_1 layers is seen to follow a regular diurnal cycle, being maximum at noon and tapering off at both earlier and later hours as shown. The density in summer is about one-third greater than in winter. This behavior of the E and F_1 layers can be attributed to ionizing radiations from the sun. The maximum electron density in the F_2 layer shows a much larger diurnal and seasonal variation, changes more from day to day, and is apparently less directly connected with the sun. It will be noted that all the maximum electron densities decrease greatly during the night, presumably as a result of recombination in the absence of ionizing solar radiation.

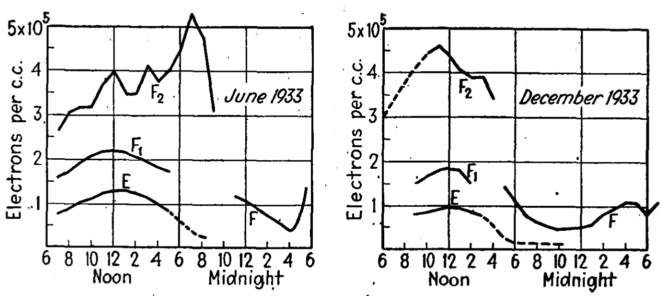


FIG. 337.—Average electron density of the various ionized layers at Washington, D. C., during 1933–1934 (quiet part of sun-spot cycle).

The situation summarized in Fig. 337 represents average results taken at Washington, D. C. during 1933–1934 and are typical of what can be expected in temperate latitudes during the quiet part of the sunspot cycle. Experiments made in Peru indicate that the latitude has little effect on the E, F_1, and F layers, but influences the characteristics of the F_2 layer noticeably.

The maximum electron density apparently follows the eleven-year sun-spot cycle, and evidence indicates that at the next sun-spot maximum

(which will be 1938–1939) the maximum electron densities will increase from 50 per cent to 100 per cent above those shown in Fig. 337, although the layer heights to be expected will not be materially different.

In addition to the E, F, F_1, and F_2 layers, there is experimental evidence to indicate the existence of other layers or maxima under certain conditions, notably a C layer lower than the E layer, and a G layer which is above the F_2 layer.[1]

118. Effect of the Ionosphere upon a Radio Wave.—The effect that an ionized region has on a radio wave can be understood by considering the behavior of a single ion or electron when under the influence of a passing radio wave. Take first the case of an electron in a vacuum with no magnetic field present other than the weak magnetic field of the wave. The wave's electrostatic field exerts forces on the electron which vary sinusoidally with time and cause the electron to vibrate sinusoidally along a path parallel with the flux lines of the wave. The amplitude and average velocity of vibration are greater the lower the frequency, and the velocity is 90° out of phase with the electric field of the radio wave because the moving electron offers an inertia reactance to the forces acting upon it. Since a moving charge is an electrical current, the vibrating electron acts as a small antenna which abstracts energy from the radio waves and then reradiates this energy in a different phase, in exactly the same manner as a parasitic antenna tuned to offer an inductive reactance (see Sec. 135). The net effect of the ionized region under these conditions is to alter the direction in which the resultant energy flows in such a manner as to cause the wave path to bend away from the regions of high electron density toward regions of lower density. The magnitude of this effect varies with the amplitude and average velocity of the electron vibrations and therefore becomes increasingly great as the wave frequency is lowered. Ions in the path of a wave act in much the same way as electrons, but because of their heavier mass ions move much more slowly than electrons under the same force and so in comparison have negligible effect.

Effect of the Earth's Magnetic Field.—The effect that electrons in the atmosphere have on radio waves is influenced by the fact that these electrons are in the presence of the earth's magnetic field, which exerts a deflecting force on the moving electron as explained in Sec. 24. The magnitude of this deflecting force is proportional to the instantaneous velocity of the electron and is in a direction at right angles to the lines of magnetic flux. At very high radio frequencies, where the maximum velocity attained by an electron in the path of the wave is small, the deflecting forces that are exerted by the earth's magnetic field are cor-

[1] Thus see Kirby and Judson, *loc. cit.*; R. C. Colwell and A. W. Friend, The Lower Ionosphere, *Phys. Rev.*, vol. 50, p. 632, Oct. 1, 1936.

respondingly slight, with the result that the path followed by the vibrating electron is a narrow ellipse (see Fig. 338*a*). The ratio of minor to major axes of the ellipse depends upon the orientation of the electrostatic flux lines of the radio wave with respect to the earth's magnetic field. The motion along the minor axis causes some of the energy reradiated by the vibrating electron to have a component polarized at 90° with respect to the polarization of the passing wave. Hence a radio wave in passing through an ionized region in the upper atmosphere has its plane of polarization affected by the earth's magnetic field.

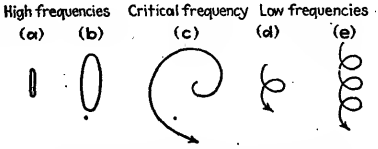


FIG. 338.—Paths followed by an electron in the earth's magnetic field when vibrating under the influence of a radio wave. To avoid confusion the paths for the low-frequency cases represent only a half cycle of the vibration, which repeats cyclically.

As the frequency of the wave is reduced, the maximum velocity attained by an electron under the influence of the wave is greater, and the minor axis of the elliptical path becomes larger in proportion to the major axis (see Fig. 338*b*), thus increasing the fraction of the absorbed energy that is reradiated in a different plane of polarization. This tendency continues as the frequency is lowered until at about 1400 kc the deflecting force of the magnetic field is so related to the frequency that the direction of the electron is reversed at exactly the same instant the electrostatic flux of the passing wave changes polarity, causing the electron to follow the spiral path shown at Fig. 338*c*. Since the electron velocity in the spiral increases without limit, the electron abstracts more energy from the passing wave than is reradiated, tending to make the sky-wave attenuation at 1400 kc greater than at higher or lower frequencies. At frequencies lower than this resonant frequency of vibration the path followed by an electron vibrating in the presence of the earth's magnetic field is approximately as shown at Fig. 338*d* and *e*. The principal effect of the magnetic field under these conditions is to lower the maximum electron velocity, thus reducing the amount of energy that is absorbed from a low-frequency radio wave and reradiated. The effect that an electron path such as shown at Fig. 338*d* and *e* has on the polarization is relatively small since the radiations from the two sides of each reversal loop are of opposite phase and tend to cancel.

A quantitative analysis shows that the presence of the magnetic field, in addition to affecting the polarization, also causes the wave to be split into two components termed the *ordinary* and *extraordinary* rays, which follow different paths, have different phase velocities, and suffer different attenuations.[1] As a consequence a wave that has

[1] A more complete discussion of the effect of the magnetic field is to be found in the literature. Thus see H. W. Nichols and J. C. Schelling, Propagation of Electric

passed through an ionized region will have both vertically and horizontally polarized components irrespective of the polarization of the waves radiated from the transmitter, and furthermore the vertical and horizontal polarizations will in general not be in the same phase, *i.e.*, the wave will be elliptically polarized.[1]

Losses in the Ionosphere.[2]—The electrons set in vibration by a passing wave will from time to time collide with gas molecules. In such a collision the kinetic energy that the electron has acquired from the radio wave is partly transferred to the gas molecule and partly radiated in the form of a disordered radio wave which contributes nothing to the transmission. The net result is therefore an absorption of energy from the passing wave. The amount of energy thus absorbed depends upon the gas pressure (*i.e.*, upon the likelihood of a vibrating electron colliding with a gas molecule) and upon the velocity that the electron acquires in its vibration (*i.e.*, upon the energy lost per collision). As a consequence most of the absorption suffered by a wave passing through the ionosphere takes place at the lower edge of the ionized region where the atmospheric pressure is greatest, and the absorption taking place high up in the layer is relatively small. Other things being equal, the absorp-

---

Waves over the Earth, *Bell System Tech. Jour.*, vol. 4, p. 215, April, 1925; Kirby, Berkner, and Stuart, *loc. cit.;* E. V. Appleton, *Jour. I.E.E. (London)*, vol. 71, p. 642, 1932.

[1] Elliptical polarization results whenever there are vertical and horizontal components having different phase. Elliptical polarization is characterized by the fact that the resulting magnetic and electrostatic fields produced by an alternating wave never at any instant pass through zero; rather the resulting fields rotate in the plane of the wave front at a rate corresponding to the frequency of the wave, while at the same time pulsating in amplitude. The resulting field produced by elliptical polarization can therefore be represented by a rotating vector of varying length as illustrated in Fig. 339. The field can never be zero because the vertical and horizontal components do not pass through zero at the same instant.

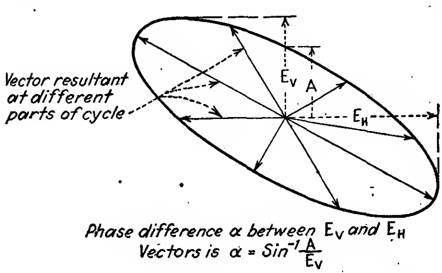


FIG. 339.—Diagram showing electric forces of an elliptically polarized wave, when the wave front is in the plane of the paper.

The sense of the polarization depends upon the direction of the earth's magnetic field, and is accordingly different in the northern and southern hemispheres. See E. V. Appleton and J. A. Ratcliffe, On a Method of Determining the State of Polarization of Downcoming Waves, *Proc. Royal Soc.*, vol. 117A, p. 576, March, 1928; A. L. Green, The Polarization of Sky Waves in the Southern Atmosphere, *Proc. I.R.E.*, vol. 22, p. 324, March, 1934.

[2] For further information on losses, see Pedersen, *loc. cit.;* William G. Baker and Chester W. Rice, Refraction of Short Waves in the Upper Atmosphere, *Trans. A.I.E.E.*, vol. 45, p. 302, 1926.

tion will also be less the higher the frequency because the average velocity of the vibrating electrons decreases as the frequency is increased.

The presence of the earth's magnetic field influences the absorption by affecting the velocity of vibration. This is particularly important at the lower frequencies, where, as explained in connection with Fig. 338, the magnetic field reduces greatly the average velocity of the electrons, and so reduces the losses.

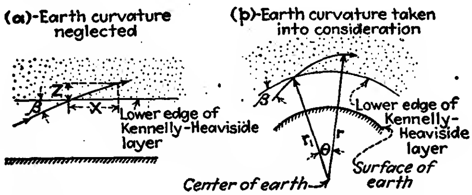


FIG. 340.—Diagrams illustrating notation of Eqs. (209a) and (209b).

119. Path Followed by Sky Wave.—The actual path that a radio wave follows in the ionosphere depends upon the way in which the refractive index of the ionized medium varies with height above the earth. When the refractive index is a function only of the height above earth, as is normally the case, the path is determined by the following differential equations:[1]

For flat earth (for short distances along earth's surface):

$$dx = \frac{\cos \beta dz}{\sqrt{\mu^2 - \cos^2 \beta}} \tag{209a}$$

For round earth (for great distances along earth's surface):

$$d\theta = \frac{r_1 \cos \beta dr}{r\sqrt{r^2\mu^2 - r_1^2 \cos^2 \beta}} \tag{209b}$$

where

β = angle of incidence of wave entering the Kennelly-Heaviside layer
μ = refractive index of Kennelly-Heaviside layer, and is a function only of z
z = height of ray above lower edge of layer
x = distance parallel to earth's surface, measured from the point at which the wave enters the ionized region
θ = polar angle (in radians) intercepted by wave path
r = distance from wave to center of earth
r_1 = distance from lower edge of Kennelly-Heaviside layer to center of earth.

This notation is shown graphically at Fig. 340.

[1] The derivation of these equations is to be found in a number of places in the literature. Thus see Baker and Rice, *loc. cit.*

The refractive index μ takes into account the effect that the electrons have on the wave path. It is determined primarily by the frequency of the wave and the electron density, although it is affected somewhat by the earth's magnetic field and by collisions between the vibrating electrons and gas molecules. If the effects produced by the magnetic field and the collisions are neglected, the refractive index is[1]

$$\text{Refractive index } \mu = \sqrt{1 - \frac{81N}{f^2}} \tag{210}$$

where

N = electron density in electrons per cubic centimeter

f = frequency in kilocycles.

It will be observed that the extent to which the refractive index departs from the value of unity that exists in free space, and hence the effect that the ionization produces on the wave, increases directly with the electron density N and inversely as the square of the frequency. When the effect of the energy loss in the ionosphere is taken into account, it is found that the refractive index is slightly closer to the free space value of unity than given by Eq. (210). The presence of a magnetic field likewise modifies the refractive index, and also leads to two values of refractive index corresponding to the two wave paths mentioned above as being produced by a magnetic field.[2]

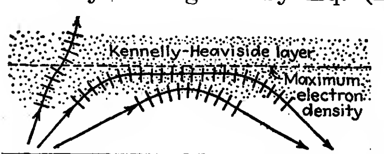


FIG. 341.—Diagram illustrating the change that takes place in the direction of the wave front when a wave travels through the Kennelly-Heaviside layer. The part of the wave-front in the region of greatest electron density advances faster than the rest of the wave and so causes a refraction.

It can be shown from Eq. (209) that the effect of an ionized region is to bend the path of the wave away from the regions of low refractive index (high electron density) toward regions of higher refractive index (low electron density). The curvature is proportional to the rate of change of refractive index with height, and so is greatest when the electron density changes rapidly with height.

A physical picture of the situation as it affects the sky wave can be gained with the aid of Fig. 341, which shows small sections of wave fronts traveling in different parts of an ionized region having a single maximum of ionization. The edge of the wave front where the electron density is greatest has the lowest refractive index and hence the highest phase velocity. This part of the wave front hence advances faster than the rest, causing a bending of the wave path away from the region of maxi-

[1] For derivation see Pedersen, *loc. cit.;* or Baker and Rice, *loc. cit.*

[2] Formulas for the refractive index in the presence of a magnetic field are given by Kirby, Berkner, and Stuart, or Appleton, *loc. cit.*

mum ionization.[1] When the wave enters the ionized region at glancing incidence, as illustrated by the lower path in Fig. 341, it needs to be bent through only a relatively small angle to be returned to earth, and it is able to obtain the necessary curvature without penetrating very deeply into the layer. In the case of a greater angle of incidence, shown by the middle path of Fig. 341, the wave penetrates more deeply into the layer and almost but not quite reaches the region of maximum density. At these higher parts of the path the variation of electron density with height is relatively small, so that the curvature is slight and the wave must travel a long distance in this higher part of the region before its path is sufficiently bent to bring the wave back into the lower regions. In the case of almost vertical incidence, as shown by the left-hand path in Fig. 341, the ionized region is not able to produce sufficient curvature to return the wave to earth before it reaches the region of maximum electron density. The wave then passes on through the layer and is lost unless there is a higher region where the maximum electron density is greater than in the layer shown.

Conditions Required to Refract Wave to Earth.—By combining Eq. (210) with Snell's law it is possible to deduce the conditions under which a wave entering the ionosphere will be returned to earth, without actually calculating the entire path that the wave follows.[2] According to Snell's law, when a wave passes from air or free space into a refracting medium, the relationship between the angle of incidence of the wave β and the angle of refraction α is given by the equation

$$\cos \beta = \mu \cos \alpha \tag{212}$$

[1] It is well at this point to note the distinction between phase velocity on the one hand and signal or group velocity on the other. The phase velocity is determined by the rapidity with which the phase changes along the path of a wave and is equal to the velocity of light divided by the refractive index. In the case of waves traveling in the Kennelly-Heaviside layer the phase velocity is hence greater than the velocity of light and may even approach infinity under some conditions.

The signal or group velocity represents the velocity with which the energy of the wave travels rather than the rate at which the phase changes. The signal velocity cannot exceed the velocity of light and may even approach zero in certain limiting cases. The phase and signal velocities are related to each other by the equation

$$\frac{\text{Group velocity}}{\text{Phase velocity}} = \frac{1}{1 - \dfrac{\omega dv}{v d\omega}} \tag{211}$$

where $\omega = 2\pi f$ and v is phase velocity. See Pedersen, *op. cit.*, p. 169.

[2] Actual calculations of wave paths cannot be carried out with any degree of certainty because of the lack of knowledge concerning the distribution of the electron density with height in the ionosphere. In calculating wave paths it is therefore necessary to make some hypothetical assumptions as to the distribution. An example of such a calculation is given by G. W. Kenrick and C. K. Jen, Measurements of the Height of the Kennelly-Heaviside Layer, *Proc. I.R.E.*, vol. 17, p. 711, April, 1929.

where α is the angle of refraction and the refractive index is μ. This notation is illustrated in Fig. 342. Furthermore it is to be observed that, if the wave is to be bent back to earth, then at its highest point $\alpha = 0$, and the refractive index at the apex of the path is

$$\mu = \cos \beta \tag{213}$$

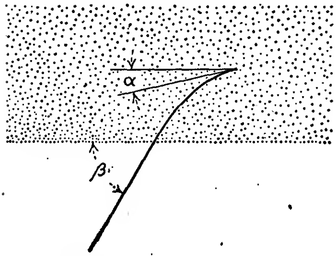


FIG. 342.—Diagram illustrating notation of Eq. (212).

By substituting into Eq. (213) a mathematical expression for the refractive index, one obtains the relation between the frequency and the maximum electron density that the ionosphere must have in order barely to return to earth waves entering the ionized region when the angle of incidence is β. Thus, if the effect of the losses and the earth's magnetic field are neglected, the combination of Eqs. (210) and (213) gives

$$\cos \beta_0 = \sqrt{1 - \frac{81N}{f^2}} \tag{214}$$

where β_0 is the largest vertical angle for which the wave is bent back to earth. The consequences of Eq. (214) are shown in Fig. 343 for a typical case.

When the ionosphere has two or more layers (*i.e.*, two or more regions of maximum electron density) as shown in Fig. 335, the situation becomes more complicated. The behavior in a typical case is illustrated in Fig. 344. Waves in passing into the lower layer tend to be refracted back to the earth unless they penetrate sufficiently to reach the region of maximum electron density. If the electron density, angle of incidence, and frequency are such that the waves will not penetrate to the center of this lower layer, they will be bent back to the earth along a path such as *a* of Fig. 344. If, however, the maximum electron density in the lower layer is insufficient to bend the path back to earth, then the wave will pass on through the first layer and into the second layer. If the maximum electron density in this second layer is sufficiently greater than the electron density in the first layer to refract the wave to earth, the wave will follow the path *b*, whereas, if the

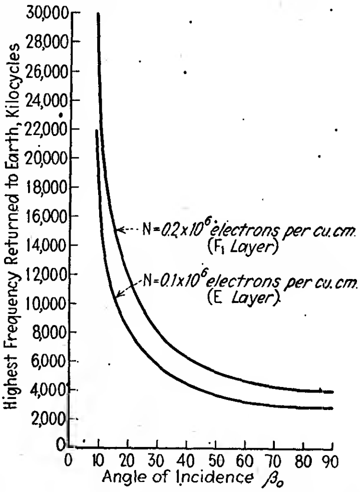


FIG. 343.—Highest frequency that can be refracted to earth plotted as a function of the angle of incidence of wave at the edge of the inosphere.

maximum electron density is insufficient to return the wave, it will pass on through as shown by path c and will never return unless there is a still higher layer of greater electron density.

As a result of this situation, the path that is followed by a wave that returns to earth will depend upon the frequency and the angle of inci-

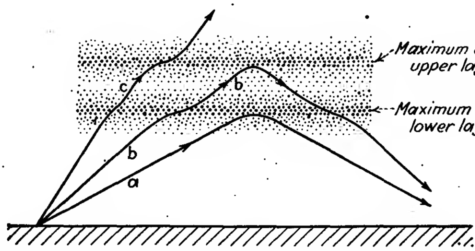


FIG. 344.—Wave paths in two-layer ionosphere for different angles of incidence.

dence. Thus, if two layers have electron densities of 0.2 and 0.1 × 10^6 electrons per cubic centimeter, reference to Fig. 343 shows that, for a frequency of 5000 kc, the lower or E layer will return the wave to earth when the angle of incidence does not exceed 35 deg., while the upper or F_1 layer returns those waves having an angle of incidence between 35 and 54 deg. For angles greater than 54 deg. this frequency is not returned to earth by either layer.

In calculating the behavior of a wave in the ionosphere it is to be noted that with small angles of incidence the angle between the wave path and the surface of the earth differs from the angle of incidence at the ionosphere because of the curvature to the earth. The relationship can be readily shown to be

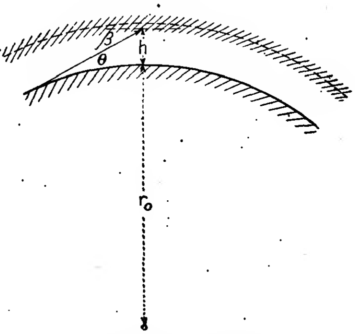


FIG. 345.—Diagram illustrating notation of Eq. (215).

$$\cos \beta = \frac{\cos \theta}{1 + \dfrac{h}{r_0}} \tag{215}$$

where

β and θ = the angles illustrated in Fig. 345

h = height of the lower edge of the ionosphere above the earth

r_0 = radius of the earth.

The angles β and θ will be virtually the same except for glancing incidence. In the limit where θ is zero, it will be found that the smallest possible angle of incidence at the ionosphere is of the order of 10 to 15 deg., with the exact value depending upon the height of the layer above earth.

Reflection of Wave by Ionosphere.—In the above discussion it has been tacitly assumed that the change in refractive index in a distance corresponding to one wave length is negligibly small. If this is not true, reflection as well as refraction takes place. At the lowest radio frequencies, where the wave length is long, considerable reflection takes place, causing the ionosphere to act as a mirror into which the waves penetrate only a small distance. At high frequencies a small amount of reflection may take place at the E layer, particularly when the refractive index is such that the wave is just barely able to penetrate the E layer.[1]

Reflection of Sky Wave by the Ground.[2]—When a sky wave that has been refracted earthward by the ionosphere strikes the ground, it is reflected. The angle of reflection is equal to the angle of incidence, but because of the fact that the earth is not a perfect conductor the reflection is incomplete. The exact behavior depends upon the angle of incidence, the polarization of the wave, the frequency, and the properties of the earth, and it can be treated as an optical reflection. With *plane waves* the vector ratio of reflected wave to incident wave is:

Component of vertical polarization (*Fig.* 346*a*):

$$\frac{\text{Reflected wave}}{\text{Incident wave}} = \frac{\epsilon \sin \delta - \sqrt{\epsilon - \cos^2 \delta}}{\epsilon \sin \delta + \sqrt{\epsilon - \cos^2 \delta}} \tag{216a}$$

Component of horizontal polarization (*Fig.* 346*b*):

$$\frac{\text{Reflected wave}}{\text{Incident wave}} = -\frac{\sqrt{\epsilon - \cos^2 \delta} - \sin \delta}{\sqrt{\epsilon - \cos^2 \delta} + \sin \delta} \tag{216b}$$

where

δ = angle of incidence as in Fig. 346
$\epsilon = k - j6\sigma\lambda \times 10^{10}$
k = dielectric constant of earth (e.s.u.)
σ = earth conductivity (e.m.u.)
λ = wave length in centimeters
$j = \sqrt{-1}$.

The quantity on the right-hand side of Eqs. (216*a*) and (216*b*) represents the reflection coefficient. The absolute magnitude of this coefficient

[1] For further discussion of reflection, see T. R. Gilliland, G. W. Kenrick, and K. A. Norton, Investigation of Kennelly-Heaviside Layer Heights for Frequencies between 1600 and 8650 Kilocycles per Second, *Proc. I.R.E.*, vol. 20, p. 286, February, 1932; Shogo Namba, General Theory of the Propagation of Radio Waves on the Ionized Layer of the Upper Atmosphere, *Proc. I.R.E.*, vol. 21, p. 238, February, 1933.

[2] For further discussion see C. B. Feldman, The Optical Behavior of the Ground for Short Radio Waves, *Proc. I.R.E.*, vol. 21, p. 764, June, 1933; G. W. O. Howe, Reflection of Waves at Earth's Surface, *Wireless Eng. and Exp. Wireless*, vol. 11, p. 59, February, 1934.

gives the ratio of reflected wave amplitude to incident wave amplitude, while the angle of the reflection coefficient is the amount by which the phase of the reflected wave differs from the phase that would exist for a perfect earth. The phase for the latter case is illustrated in Fig. 346*a* and *b*.

The phase and magnitude of the reflection coefficient as a function of angle of incidence are shown in Fig. 346*c* for a typical short-wave case. It will be noted that for vertical polarization there is a particular angle at which the reflection coefficient passes through a minimum, and

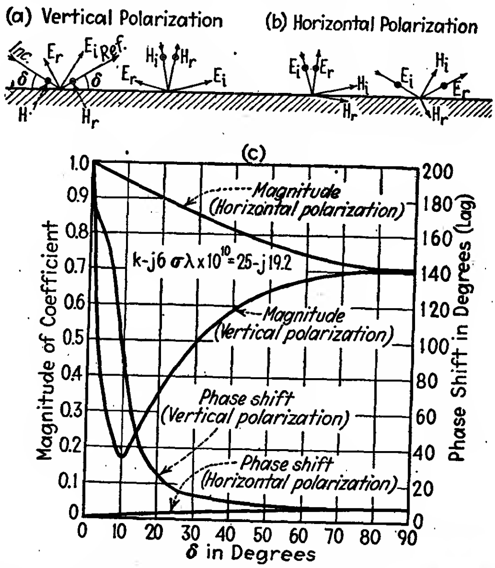


FIG. 346.—Reflection coefficient as a function of angle of incidence, together with diagram illustrating assumed positive polarities for Eq. (216) for the case of imperfect earth. (E and H denote electric and magnetic flux, respectively, and subscripts i and r denote incident and reflected components.)

that for more glancing incidence the phase angle of the reflected wave approaches 180 deg. (*i.e.*, is reversed). This angle of incidence at which the reflection coefficient is minimum corresponds to the Brewster angle encountered in optics and decreases as the earth conductivity is increased.

120. Sky-wave Attenuation.[1]—The sky wave in traveling between the earth and the lower edge of the ionosphere suffers practically no attenuation except that due to spreading, so that in these regions the field strength is inversely proportional to distance. In the ionosphere, however, energy is absorbed from the wave as a result of collisions between the vibrating electrons and gas molecules as explained in Sec. 118. This absorption of energy causes the ionized region to have an

[1] For further discussion see Namba, *loc. cit.*; or Pedersen, *loc. cit.*

equivalent conductivity which depends upon the frequency of collisions and the electron density and which results in an added attenuation. The lack of knowledge concerning the exact nature of the upper atmosphere, particularly the distribution of electrons and the variation of gas pressure with height, makes it impossible to predict accurately by calculation the strength of the sky wave that is refracted back to earth. Theoretical calculations of the attenuation produced by the ionosphere under hypothetical conditions have, however, been found to be of considerable value in explaining observed phenomena. Thus, when a wave penetrates through an ionized layer as in the case of paths b and c in the lower layer of Fig. 344, the attenuation in this layer is inversely proportional to the square of the frequency, and becomes smaller the more nearly vertical the angle of incidence. However, when the wave is refracted sufficiently to return to earth, as path a in the lower layer of Fig. 344, the attenuation in this layer depends upon the way in which the electrons are distributed with height, and becomes greater as the angle of incidence at the layer approaches vertical.

When a wave passes an appreciable distance into the ionosphere, most of the attenuation that results occurs at the lower edge of the ionized region, which means either in the E layer for waves that penetrate to higher layers or at the lower edge of the E layer for waves returned to earth by the E layer. This is because it is here that the atmospheric pressure is greatest. The attenuation high up in the ionized region is comparatively small because, although the electron density is higher, the atmospheric pressure is very low at the high elevation. Inasmuch as the energy lost in passing through a section of the ionosphere is inversely proportional to the square of the frequency, it is to be noted that the loss in the lower edge of the ionized region and hence the attenuation tend to be less the higher the frequency.

121. Propagation of Low-frequency Radio Waves (20 to 550 Kc).—The propagation of low-frequency waves is characterized by a relatively low ground-wave attenuation and by the fact that the sky wave is refracted back to earth after only a very slight penetration into the ionosphere and with little absorption. The low ground-wave absorption at low frequencies is to be expected from the Sommerfeld analysis, and is so marked that at the lowest frequencies used in communication most of the energy reaching receiving points up to 1000 km travels by way of the ground wave at all times. Under such conditions the received signal is very nearly independent of diurnal, seasonal, and yearly variations.

At great distances from the transmitter the ground wave is almost completely absorbed, and most of the received energy represents sky wave. Because the ionosphere refracts the low frequencies so readily, the action at considerable distances from the transmitter is practically

the same as though the wave were propagated in the space between two concentric reflecting spherical shells representing the earth and the lower edge of the ionosphere.[1] The attenuation under such conditions is that caused by spreading, plus the added loss occurring at the earth's surface and the edge of the ionosphere. The loss in the ionosphere is normally less than at the earth's surface, but varies with the conditions at the lower edge of the ionized region and so has a diurnal and a seasonal variation. The loss also follows the eleven-year sun-spot cycle. As a consequence the signals received at considerable distance are variable, being weakest when the path between transmitter and receiver is in daylight and being generally stronger in winter than in summer. Transmission is also characterized by a sudden drop in signal intensity when the sunset line falls across the transmission path. Typical curves showing diurnal variation in field strength of long-distance long-wave signals at different times of year are shown in Fig. 347.[2] The principal seasonal effect is longer night and shorter day transmission conditions in winter. The seasonal and diurnal variations become more pronounced as the frequency is raised, as is brought out by Fig. 347. While low-frequency radio signals normally behave in a fairly regular manner, neither the daily nor yearly cycles repeat exactly. There is also evidence to show that low-frequency radio waves propagate in a north-south direction with less attenuation than when traveling in an east-west direction.[3]

[1] Under these conditions the daytime field strength when the transmission path is over water is given with fair accuracy by the revised Austin-Cohen formula, which is as follows:

$$\epsilon = 377\frac{hI}{\lambda d}\sqrt{\frac{\theta}{\sin\theta}}\,e^{\frac{-0.0014d}{\lambda^{0.6}}} \times 10^3 \tag{217}$$

where

e = base of Napierian logarithms
ϵ = field strength in microvolts per meter
h = effective height of transmitting antenna in kilometers
λ = wave length in kilometers
d = distance in kilometers to transmitter
θ = angle at center of earth intercepted by transmission path in radians
I = current flowing in the vertical part of the transmitting antenna.

This formula was derived empirically, but has been found to have some theoretical justification. See L. W. Austin, Preliminary Note on Proposed Changes on the Constants of the Austin-Cohen Transmission Formula, *Proc. I.R.E.*, vol. 14, p. 377, June, 1926; G. N. Watson, The Transmission of Electric Waves around the Earth, *Proc. Roy. Soc.*, A, vol. 95, p. 546, July 15, 1919.

[2] These curves are taken from Lloyd Espenschied, C. N. Anderson, and Austin Bailey, Transatlantic Radio Telephone Transmission, *Proc. I.R.E.*, vol. 14, p. 7, February, 1926.

[3] See Eitaro Yokoyama and Tomozo Nakai, East-west and North-south Attenuations of Long Radio Waves on the Pacific, *Proc. I.R.E.*, vol. 17, p. 1240, July, 1929.

At intermediate distances the sky and ground waves are both of appreciable magnitude, causing the observed field strength to be the vector sum of two waves which have traveled over paths of unequal lengths and so in general are not of the same phase. Since this phase difference depends upon the distance to the transmitter, there is tendency for the field strength to increase and decrease alternately as the distance from receiver to the transmitter is increased.[1] The relative phase of the two component waves at any particular location depends upon the height of the Kennelly-Heaviside layer as well as the distance to the transmitter, and so shows diurnal and seasonal variations. The result is that, while in summer the signals received from a not-too-

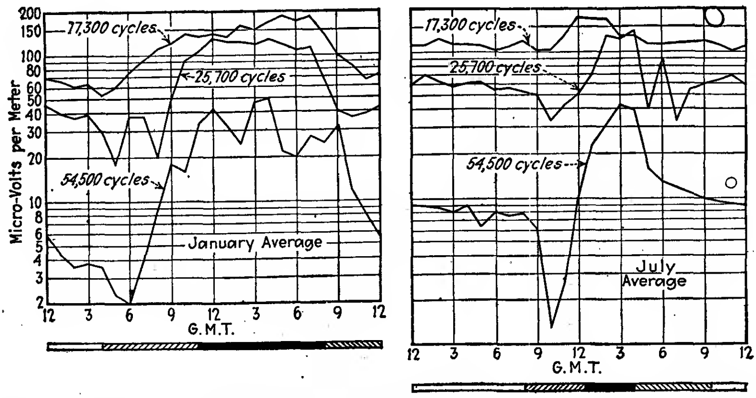


FIG. 347.—Curves showing average diurnal variation in strength of long-wave signals of different frequencies propagated across the north Atlantic during mid-winter and midsummer months. (Note that signal strengths at different frequencies cannot be compared because the radiated power was not the same at all frequencies.) The solid and clear strips at the bottom of the figure indicate periods when the entire transmission path is in darkness and light, respectively, while the shaded strips indicate part of the path in darkness and part in light.

distant transmitter may be stronger during the day than at night, the opposite may be true in winter, while at another receiving location these actions may be entirely reversed.

A physical picture of the more important phenomena involved in long-wave propagation may be obtained by considering the various paths by which energy may reach a receiving point, as illustrated in Fig. 348. The sky-wave paths penetrate only slightly into the ionized region, which acts very much as a mirror, turning back the waves with an angle of reflection equal to the angle of incidence. When these waves strike the earth they are likewise reflected, with the angle of reflection equal to the angle of incidence. Each time such a reflection or refraction of

[1] See J. Hollingworth, Propagation of Radio Waves, *Jour. I.E.E.* (*London*), vol. 64, p. 579, May, 1926.

the sky wave takes place there is a certain amount of absorption of energy, so that, although energy may theoretically reach a distant receiving point either by following long-step or short-step paths, most of the energy actually arrives via long steps because the other paths suffer greater attenuation.

In view of the fact that the long-step paths represent energy radiated at a low angle with respect to the earth's surface, as also does the ground wave, it is apparent that the antenna radiating low-frequency energy should concentrate this energy as far as is practical at low vertical angles.

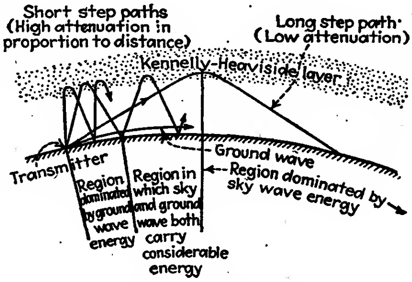


FIG. 348.—Schematic diagram illustrating paths followed by energy radiated from a long-wave transmitting antenna.

Low-frequency radio waves are always vertically polarized when observed near the earth's surface and normally have a wave front that is tilted somewhat forward. The tilt is caused by earth losses while the vertical polarization is a result of the fact that the earth reflections cancel the horizontal component of polarization near the ground.

Fading, *i.e.*, rapid change in signal strength, is seldom observed at low frequencies except occasionally during the sunset period or at night, and even then the fading is relatively slow compared with fading of broadcast and higher frequencies.

122. Propagation of Waves of Broadcast Frequencies (Frequency Range 550 to 1500 Kc).—At broadcast frequencies the ground-wave attenuation is appreciably greater than at lower frequencies, and increases rapidly as the high-frequency end of the band is approached. Sky waves of all broadcast frequencies are practically completely attenuated during the day, but the night attenuation of the sky wave varies greatly with the conditions in the ionosphere, sometimes being moderately high, and at other times, particularly during the winter nights, becoming almost zero. The result of this situation is that the daytime coverage of broadcast stations is controlled by the Sommerfeld ground wave. The daytime range is therefore small, but the signals can always be depended upon to be the same day after day, winter and summer, and year after year. At night the Sommerfeld ground wave still accounts for most of the energy received at near-by points, but as a result of the sky wave appreciable energy reaches very distant points which are far beyond the range of the daytime signals.

The situation that exists at night is illustrated in detail in Fig. 349, where it is seen that there are three distinctive zones. Near the trans-

mitter the sky wave is relatively weak compared with the ground wave and the latter predominates. As the distance to the transmitter is increased, the ground wave becomes attenuated, whereas the sky wave becomes stronger, thus making the ground and sky waves of approximately equal strength. At still greater distances the sky wave tends to become still stronger and to maintain a relatively high and constant signal strength up to considerable distances.

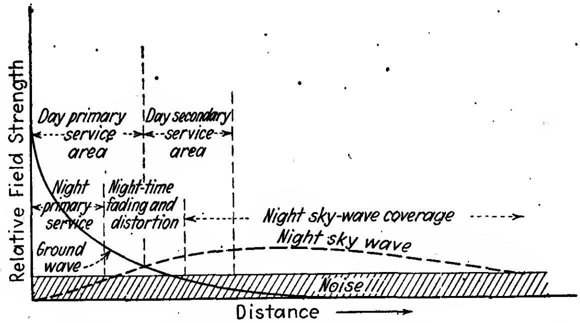


FIG. 349.—Diagram illustrating different types of coverage obtained from a high-power broadcast station during the day and night periods. The shaded area represents the lowest field strength which completely overrides the noise level.

The reason for this behavior of the sky wave lies in the fact that, as the distance from the transmitter is increased, the sky wave that reaches the receiver represents energy radiated from increasingly lower vertical angles, as illustrated in Fig. 350, and the characteristics of broadcast antennas are such that the radiated energy is greater the lower the angle above the horizon. Furthermore, in view of the fact that the sky wave reaches a height of about 100 km at broadcast frequencies, the sky wave

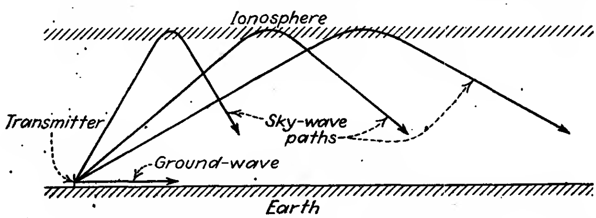


FIG. 350.—Paths by which a broadcast station delivers energy at various distances.

must travel nearly as far to reach receivers near the transmitter as receivers several hundred kilometers away. The sky-wave attenuation and path at broadcast frequencies are to a first approximation independent of frequency, so that the distant night-time reception, which depends upon the sky wave, is about equally satisfactory for all frequencies. This is in contrast with the daytime (or ground-wave coverage) for which the Sommerfeld analysis shows low frequencies are the most effective.

The middle night zone illustrated in Fig. 349, where the sky wave and ground wave are of approximately equal intensity, is of particular importance since it represents a region where there is severe fading and bad quality at night. Here the night signals represent the vector sum of two waves that have traveled along different paths. At some places these waves will add, and at other locations they will subtract, resulting in the formation of an interference pattern over the surface of the earth in this region. This pattern is unstable, since very slight changes in the conditions existing in the ionosphere will be sufficient to change the length of the sky-wave path by half a wave length and thereby change the interference from an addition to a subtraction, or vice versa. The result is that in this region, where the ground and sky waves are of approximately equal strength, severe fading is always experienced.

Furthermore the interference pattern is very sensitive to frequency because a slight change in frequency will change the relative path lengths by an appreciable fraction of a wave length. Inasmuch as a modulated wave consists of a carrier and a series of side-band frequencies, the different parts of the modulated wave will be treated differently, some frequency components tending to cancel while others tend to add in the interference pattern. Investigations of broadcast signals have shown that frequencies differing by as little as 250 cycles often fade in and out independently of each other.[1] This phenomenon is termed *selective fading*, and results in a loss in quality which is commonly so severe as to destroy completely the entertainment value of the broadcast signal.

At great distances there is ordinarily a certain amount of night-time fading and quality distortion, presumably as a result of two or more sky waves reaching the receiver along different paths. However, when only sky waves are involved, the effect is much less severe than when both sky and ground waves are involved.

Broadcast Coverage.—The primary object of most broadcast transmitters is to deliver a high-quality signal continuously to listeners within a limited distance of the transmitter rather than to produce an understandable signal at great distances, as is the case with most other types of radio communication. The region around the transmitter in which reception free from objectionable disturbances or distortion can be obtained at all times is called the primary service area of the broadcast station and represents the region in which the ground wave is powerful enough to override all ordinary interference (either natural or man made) as well as the sky wave. Within the primary service area the signals

[1] A very thorough study of fading phenomena on broadcast frequencies is to be found in the classic article of Ralph Bown, De Loss K. Martin, and Ralph K. Potter, Some Studies in Radio Broadcast Transmission, *Proc. I.R.E.*, vol. 14, p. 57, February, 1926.

will show little or no seasonal and diurnal variation in strength and will have an entertainment value that is not reduced by interference. The strength of ground wave required to overcome local interference varies with the location of the receiver and is much higher in cities, particularly in industrialized districts, than in rural regions. Experience has shown that satisfactory service in metropolitan areas requires a signal strength in the order of 5 to 30 mv per meter, while in rural regions fairly satisfactory service is frequently obtained with signal strengths as low as 0.1 mv per meter.[1]

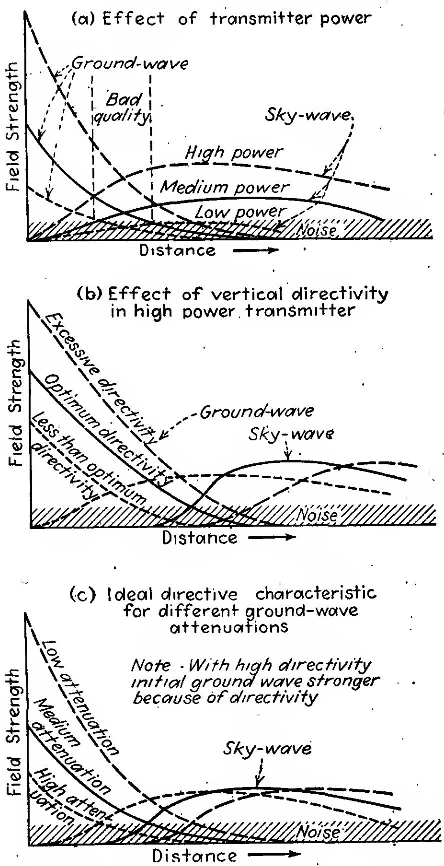


FIG. 351.—Effect of various factors on the coverage of broadcast stations.

Outside of the primary service area there is a secondary service area where the signals are sufficiently strong to give fair signals, but where the reception is not as perfect as in the primary area. Thus in the daytime secondary coverage is obtained just outside the primary service area, where the ground wave is still appreciable but is not sufficiently strong to give primary service conditions. Also at night with powerful transmitters the sky wave is sufficiently strong to give secondary coverage over large areas.

The factors that determine the coverage of the signals are the transmitter power, the directive characteristics of the transmitting antenna, the transmitter frequency, the soil characteristics, and the noise level. The relationship of these various factors to the problem can be understood with the aid of Fig. 351. Consider first the effect of the transmitter power as shown at Fig. 351*a*. With low power the situation is as shown by the dotted line. Here the primary service area is small and the sky wave is everywhere too weak to provide

[1] See Lloyd Espenschied, Radio Broadcast Covering of City Areas, *Trans. A.I.E.E.*, vol. 45, p. 1278, 1926; C. M. Jansky, Jr., Some Studies of Radio Broadcast Coverage in the Mid-west, *Proc. I.R.E.*, vol. 16, p. 1356, October, 1928.

secondary coverage, so that the secondary service area represents weak ground-wave coverage and is the same both night and day. Increasing the transmitter power produces the situation shown by the solid line, in which the primary service area has been increased and the ground wave is strong enough to provide secondary coverage out to the region of high distortion and beyond. The result is that during the daytime the secondary coverage obtained by ground-wave transmission is considerably more than at night, because at night the region of bad selective fading that the night sky wave produces lies in the daytime secondary coverage area. At night, however, the sky wave gives fair secondary coverage at very great distances. If the power is increased still further, as shown by the dashed line, it will be found that, whereas the daytime primary and secondary coverages are increased, the night-time primary service area is no larger than with less power because the outer portion of the daytime primary service area is in the night-time bad-fading region. In order to take full advantage of the increased power during the night, it is necessary to modify the directive characteristics of the antenna.

The effect upon the coverage obtained by concentrating the radiated energy in different amounts along the horizontal is shown in Fig. 351*b* for a very high-power transmitter. The increased horizontal concentration strengthens the ground wave, thereby increasing the daytime coverage. At the same time the reduction in high-angle radiation reduces the strength of the sky wave returning to earth close to the transmitter. These two actions combine to push the region of bad fading out farther from the transmitter. The increased sky wave at low vertical angles also increases the strength of the night sky wave reaching distant points.

It will be noted from Fig. 351*b* that there is an optimum amount of vertical directivity. Thus the curve in Fig. 351*b* marked "Excessive directivity" reduces the high-angle sky wave to such a degree that the ground wave virtually disappears before the sky wave returns in appreciable amount. This gives rise to a region where adequate signals are never received either during the day or during the night, although more distant points receive relatively strong sky waves at night. The ideal directional characteristic for the antenna of a high-power broadcast station would be one producing a result such as illustrated by the solid line in Fig. 351*b*, in which the sky wave is negligibly small until the strength of the ground wave becomes too weak for satisfactory secondary coverage, and then comes in abruptly with a strength somewhat greater than the ground wave. Such a night sky wave does not reduce the night coverage of the ground wave below the day coverage, and at the same time it keeps the width of the region of high distortion down to a minimum. The exact details of the ideal characteristic will depend upon

the frequency, power, noise level, soil conductivity, and conditions in the ionosphere. Increased soil conductivity and lower frequency both increase the strength of the ground wave without altering the sky wave appreciably, thereby making it desirable to have greater horizontal directivity, as shown in Fig. 351*c*. Increased power also makes increased low-angle directivity desirable, as is apparent by comparing Figs. 351*a* and 351*b*. Lowering the noise level has much the same effect as increasing the power, since the nature of the coverage is determined primarily by the ratio of signal to noise. The diurnal and seasonal variations of the height and electron density of the ionosphere affect the strength of the sky wave, and also modify the point on the earth's surface at which waves radiated in a given direction return. Accordingly, the optimum directivity is continually varying, and the best that can be done is to approximate the average situation to be expected.

If the transmitting antenna is relatively short, the field strength about the antenna is proportional to the cosine of the angle of elevation, as illustrated in Fig. 3. In order to obtain more directivity, it is necessary either to employ a tower having a height not less than half a wave length or to use a relatively complicated antenna array. Either type of directive antenna is expensive and so can be justified only in high-power installations. With low transmitter power it is much cheaper to improve the coverage by increasing the power than by increasing the directivity of the antenna system. The maximum directivity that it has been found economical to obtain is that given by a tower antenna approximately 0.53 wave length high, such as discussed in Sec. 137. This arrangement gives considerable improvement, approximately doubling the coverage of 50-kw transmitters, but is not capable of giving as much directivity as could be utilized to advantage with higher powers under conditions of low ground-wave absorption.

Examination of Fig. 351*a* shows that the increased ground-wave coverage obtained by increasing the transmitter power becomes proportionately less as the power is increased. This is because of the high ground-wave absorption when the Sommerfeld numerical distance exceeds 3 to 5, causing most of the energy radiated from the transmitter to be absorbed before the ground wave reaches such distances. The maximum power that it is practical to use for ground-wave coverage depends therefore upon the cost of the increased power and other economic matters and upon the transmitted frequency and soil conductivity. This economic power becomes greater the lower the ground-wave absorption.

The ground-wave coverage of broadcast signals is very sensitive to frequency and soil conductivity. As a result the lower broadcast frequencies are much superior to the higher frequencies when maximum

ground-wave coverage is important. Thus examination of Fig. 333 shows that it is impracticable to cover a 200-km radius with a 1-mv ground wave at 1500 kc even with good soil conductivity, although this is entirely feasible at the lower broadcast frequencies. Soil conductivity is also important, and with unfavorable conditions such as exist in some parts of the country, notably New England, even the lowest broadcast frequencies require impracticably large powers to obtain a 1-mv-per-meter ground wave at distances of only 100 km.

In certain instances sky-wave coverage at night is more important than primary service area representing ground-wave coverage. This is the case in sparsely populated regions where the only way a large number of people can receive any kind of radio service is through nighttime secondary coverage produced by powerful sky waves. In such circumstances it is desirable that the transmitter power be adequate to produce a strong sky wave, and it is desirable to concentrate the energy at low vertical angles in order to improve long-distance coverage. The frequency is of little importance, however, so that the higher broadcast frequencies, which are inherently unsuitable for ground-wave coverage, can be used.

Calculation of Broadcast Coverage.[1]—The coverage to be expected from a broadcast transmitter can usually be predicted with fair accuracy except possibly in the immediate vicinity of cities. The ground-wave field strength can be quite accurately determined with the aid of the Sommerfeld analysis or by use of curves such as given in Fig. 333, provided the transmitter power, frequency, soil conductivity, and antenna directivity are known. In the case of cities the buildings, particularly steel office buildings, make the rate of attenuation high and also unpredictable. The effect on the resulting ground-wave pattern is often marked, as illustrated in Fig. 352.

Sky-wave calculations are somewhat less certain because of the variable nature of the ionosphere. The usual assumptions made in estimating sky-wave field strength are as follows: (1) The height of the layer is 100 km (the E layer). (2) The angle of reflection at the layer is equal to the angle of incidence. (3) The refracted wave has a field strength 20 per cent of the strength of the incident wave. (4) The attenuation is independent of the angle of incidence at the layer. (5) The attenuation in the ionosphere is the same for all frequencies. (6) In addition to the ionosphere attenuation, the strength of the sky

[1] For additional information see P. P. Eckersley, The Calculation of the Service Area of Broadcast Stations, *Proc. I.R.E.*, vol. 18, p. 1160, July, 1930; T. L. Eckersly, Direct Ray Broadcast Transmission, *Proc. I.R.E.*, vol. 20, p. 1555, October, 1932; Glenn D. Gillett and Marcy Eager, Some Engineering and Economic Aspects of Radio Broadcast Coverage, *Proc. I.R.E.*, vol. 24, p. 190, February, 1936.

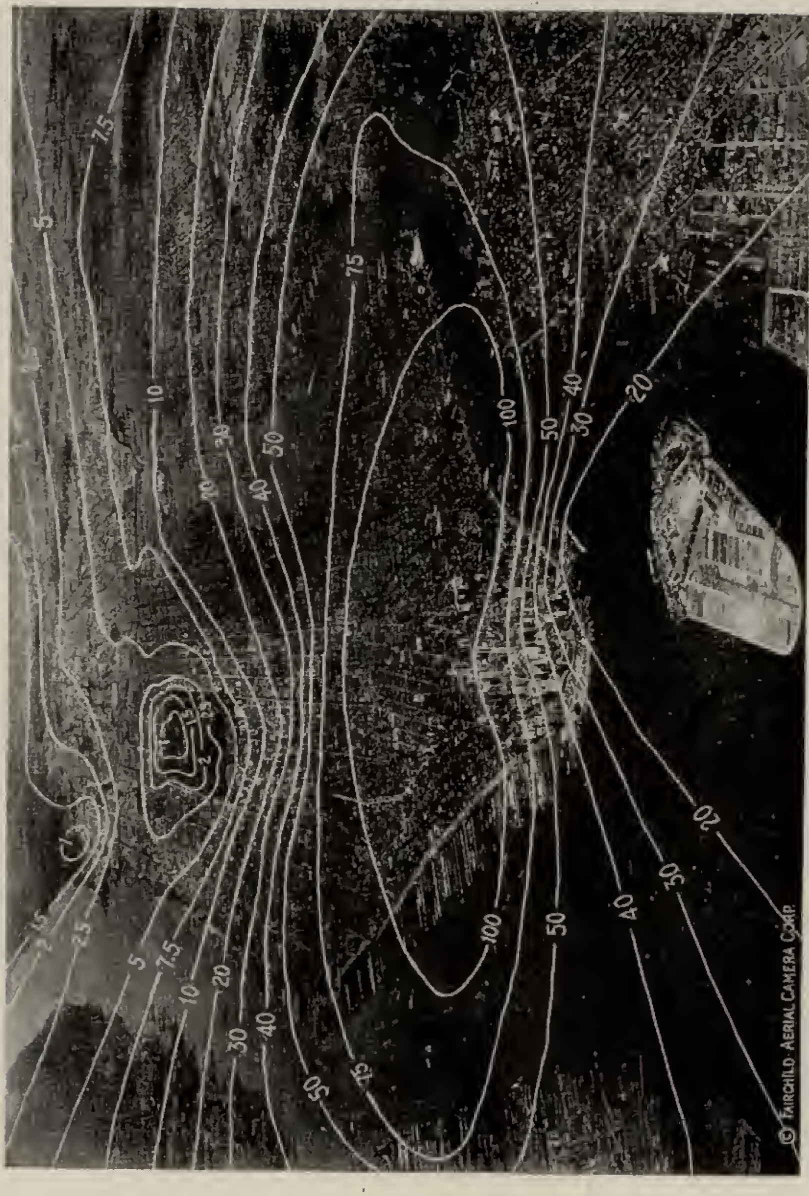


FIG. 352.—Radio field intensity in millivolts per meter for a transmitter located on the top of a building in lower Manhattan, New York City. Note the high rate of attenuation over the built-up districts, particularly the sky-scraper section, as compared with the attenuation over the water and suburban areas. (*Courtesy Fairchild Aerial Camera Corporation and the Institute of Radio Engineers.*)

wave is inversely proportional to the distance as a result of spreading. Assumptions 3 and 4 are very approximate, while assumption 5 is also open to some question as the attenuation is independent of frequency only when the electron density varies exponentially with height.[1]

In calculating the sky-wave field strength, the first step is to determine the angle above the horizontal at which energy must be radiated to return to the earth at different points, assuming that the ionosphere gives a mirror reflection. The results of such calculations are given in Fig. 353, together with the factor by which the ground-wave distance must be multiplied to give the distance covered by the sky wave. The sky-wave strength at any particular point can then be determined from a knowledge of the field strength radiated at the appropriate vertical direction and the length of the resulting sky-wave path, by assuming the attenuation is inversely proportional to distance and then taking one-fifth of the resulting strength to allow for the loss in the ionized layer.

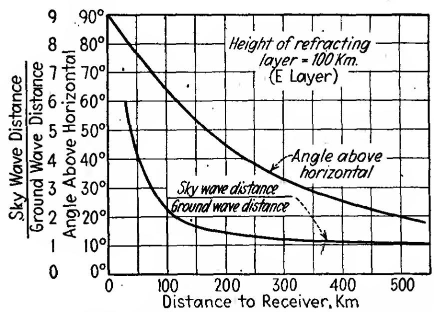


FIG. 353.—Relation between angle at which wave leaves the earth and the distance from transmitter at which return takes place, and also the factor by which the ground-wave distance must be multiplied to give the distance the sky wave travels. These curves take into account the curvature of the earth, but assume only slight penetration of the ionized *E* layer.

123. Propagation Characteristics of Short Waves (Frequency Range 1500 to 30,000 Kc).—At frequencies above 1500 kc the ground-wave attenuates so rapidly as to be of no importance except for transmission over very short distances.[2] Short-wave communication therefore ordinarily depends upon the ability of the ionosphere to refract the high-frequency sky wave back to earth at the receiving point without excessive attenuation. As a result the strength of the signal received from a distant transmitter depends on the transmitted frequency, the conditions in the ionosphere, and the angle at which the transmitted waves enter the ionized region.

[1] See Namba, *loc. cit.*

[2] The only cases where ground-wave propagation of short waves is of importance is in connection with such things as police radio, where the distances are often very small, and in moderate-distance transmission over water, where the ground-wave attenuation is very low because of the conductivity of the water. Frequencies in the range 1500 to 3500 kilocycles are used for ground-wave coverage of this type. Data on propagation of such frequencies over land, sea, and combination sea-land paths are given by C. N. Anderson, Attenuation of Overland Radio Transmission in the Frequency Range 1.5 to 3.5 Megacycles per Second, *Proc. I.R.E.*, vol. 21, p. 1447, October, 1933.

The essential features of sky-wave propagation of high-frequency waves are shown in Fig. 354, which gives hypothetical ray paths for sky waves of different frequencies leaving the transmitting antenna at different angles. In these diagrams a single ionized layer has been assumed for the sake of simplicity and, while these illustrate the general nature of short-wave propagation, they must not be taken too literally. At frequencies less than a certain critical value, waves radiated at all vertical angles are returned to earth as shown in *a*. At frequencies somewhat higher than this critical value, the waves entering the ionized region with nearly vertical incidence pass on through as shown in *b*, with the result that no sky wave returns to earth until some distance

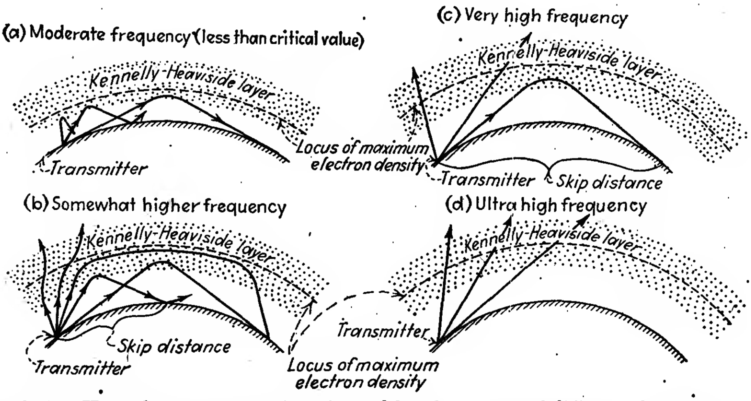


FIG. 354.—Hypothetical ray paths followed by sky waves of different frequencies. For the sake of clearness the height of the Kennelly-Heaviside layer is shown greatly exaggerated, and only a single layer is assumed.

away from the transmitter. This distance from the transmitter to where the first sky wave returns is termed the *skip distance*, and, inasmuch as the ground wave is ordinarily absorbed within a few miles of the transmitter, this skip distance represents a region where the signals are very weak or non-existent, *i.e.*, the waves skip over this territory to return at greater distances.[1] Increasing the frequency still further increases the

[1] An approximate calculation of skip distance can be made by assuming that the wave, instead of being refracted by the ionized layer, undergoes a mirrorlike reflection at the point of maximum electron density. Manipulation of the resulting geometry involving the curvature of the earth, combined with Eqs. (212) and (213), then leads to the following approximate relation giving the refractive index μ_0 that must exist at the point of maximum electron density when the skip distance is s and this maximum electron density is at a height h:

$$\mu_0^2 = \frac{\sin^2 (s/2r)}{\sin^2\left(\frac{s}{2r}\right) + \left[1 + \frac{h}{r} - \cos\left(\frac{s}{2r}\right)\right]^2} \qquad (218)$$

where r is the radius of the earth. After determining the required μ_0 for a given layer

skip distance because now only those waves entering the layer with relatively glancing incidence are returned to earth, as in c. Finally, if the frequency is made very high, even waves radiated horizontally along the earth's surface pass on through the ionosphere, as in d, and never return to earth.

When the ionosphere contains two or more layers, the situation is more involved, although the same general features are present. A typical case is illustrated in Fig. 355, where energy radiated at high angles penetrates through the lower layer and is refracted by the upper layer, whereas waves entering the ionosphere with rather glancing incidence are refracted earthward by the low layer of low electron density. There are consequently numerous paths by which waves of a given frequency might reach the receiver. A typical situation is illustrated in Fig. 356 where long-step paths such as a, b, and c and short-step paths such as d are indicated. The attenuation along these various routes differs considerably, however, so that not over three or four account for most of the received energy. This difference in attenuation arises from the fact that, as the largest part of the energy loss in the ionosphere

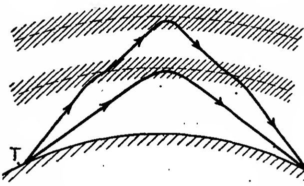


FIG. 355.—Typical sky-wave paths when two layers are present in the ionosphere.

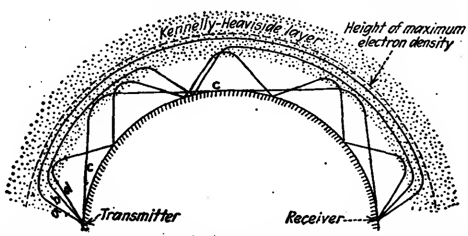


FIG. 356.—Possible routes by which short-wave signals can travel from transmitter to receiver. The height of the Kennelly-Heaviside layer has been greatly exaggerated to make the drawing clearer.

occurs at the lower edge of the ionized region, the waves that cross the absorbing region the fewest number of times (*i.e.*, the long-step paths) suffer the least absorption.

The Optimum Transmission Frequency.—Since the attenuation that a wave suffers in passing through the lower part of the ionized region is

height and skip distance, the frequency at which this skip distance is obtained can be calculated from Eq. (210) for any particular maximum electron density N in the ionosphere. For the detailed derivation see A. H. Taylor and E. O. Hulburt, The Propagation of Radio Waves over the Earth, *Phys. Rev.*, vol. 27, p. 189, February, 1926.

inversely proportional to the square of the frequency, as discussed in Sec. 120, high-frequency waves pass through the attenuating region with less loss of energy than do low-frequency waves. The optimum frequency for transmission is consequently only slightly less than the highest frequency that will return a sky wave to the receiver. A greater frequency will place the receiver inside the skip distance and so give no received signal, while a lower frequency will suffer increased attenuation.

The optimum frequency for sky-wave transmission increases with the distance to be covered and with the electron density of the layer, and is considerably greater in the daytime than at night and somewhat greater in summer than winter. For relatively short distances the frequency must not exceed the critical value for vertical incidence [see Eq. (214)], and cannot exceed the critical value appreciably even for distances up to 400 or 500 miles. In order to main continuous communication over short and moderate distances, it is usually found advisable to have available two frequencies. The lower of these is used at night and is usually in the neighborhood of 3000 kc, while the other is used in the daytime and is commonly about 6000 kc. The night frequency can be used in day as well as night, but the day signals are then weaker than when a higher day frequency is employed.

The optimum frequency for long-distance communication is approximately 20 mc in the daytime and 10 mc at night, but varies with the time of day, the season, and from year to year. In order to maintain continuous communication over large distances, it is often necessary to have available, in addition to the day and night frequencies, an intermediate or transition frequency to be used during sunrise and sunset periods and under unusual conditions when neither the day nor night frequencies are satisfactory.[1]

Difference between East-west and North-south Propagation.—Short waves propagated over great distances in an east-west direction differ markedly in their behavior from waves traveling long distances in a

[1] Data on long-distance short-wave communication, showing the character of transmission that is obtained at different hours of the day, at different seasons of the year, and at different frequencies, are given by C. R. Burrows, The Propagation of Short Radio Waves over the North Atlantic, *Proc. I.R.E.*, vol. 19, p. 1634, September, 1931; M. L. Prescott, The Diurnal and Seasonal Performance of High-frequency Radio Transmission over Various Long-distance Circuits, *Proc. I.R.E.*, vol. 18, p. 1797, November, 1930; C. R. Burrows and E. J. Howard, Short-wave Transmission to South America, *Proc. I.R.E.*, vol. 21, p. 102, January, 1933; Clifford N. Anderson, North Atlantic Ship-shore Radiotelephone Transmission during 1930 and 1931, *Proc. I.R.E.*, vol. 21, p. 81, January, 1933; Clifford N. Anderson, North Atlantic Ship-shore Radiotelephone Transmission during 1932–1933, *Proc. I.R.E.*, vol. 22, p. 1215, October, 1934; R. K. Potter and A. C. Peterson, Jr., The Reliability of Short-wave Radio Telephone Circuits, *Bell System Tech. Jour.*, vol. 15, p. 181, April, 1936.

north-south direction. This is because of the way in which the distribution of sunlight varies along the transmission path at any one time, and because of seasonal differences that exist between points on opposite sides of the equator. Thus, when communication is carried on between points having the same longitude but on opposite sides of the equator, the sunlight is more or less uniformly distributed along all parts of the path, but the seasons of the two terminals are opposite. On the other hand, the distribution of sunlight along a great-circle path lying between two points on the same latitude is non-uniform, so that one part of the path can be in sunlight while another part is in darkness. Under such a condition it is difficult to find a frequency that will propagate satisfactorily over the entire distance. Experience indicates that north-south transmission across the equator is more reliable and easier to maintain continuously than is communication over a like distance in an east-west direction. In particular, when the great-circle route between the terminals passes over the polar regions, the transmission is often unsatisfactory.

Directional Characteristics Desirable in Transmitting Antennas.—When the optimum frequency is employed and the transmission is over considerable distance, it is desirable to concentrate the radiated energy very close to the horizontal, with the best angle generally considered to be in the range 10 to 25 deg. Energy concentrated at vertical angles below about 10 deg. tends to be absorbed by the earth, and furthermore expensive antenna structures are required to concentrate the main part of the radiation at excessively low angles. The desirability of low-angle radiation for long-distance short-wave communication arises from the fact that, with the optimum frequency, the only energy that is refracted back to earth by the ionosphere is that which enters the ionized regions at almost grazing incidence, and which hence leaves the transmitter relatively close to the horizontal.

When the distance is relatively short, *i.e.*, not over several times the height of the maximum electron density in the ionosphere, the energy, in order to reach the receiver, must be radiated at an appreciable vertical angle, and marked vertical directivity is not desired.

When short-wave communication is to be carried on between two fixed points, it is desirable to concentrate the radiation in the horizontal plane as well as in the vertical, since all energy radiated in horizontal directions other than toward the receiver is wasted. By taking advantage of this possibility, the energy delivered to a distant receiver can be increased many-fold.

The amount of directivity that can be usefully employed in a transmitting antenna is limited, however, by the fact that the optimum vertical angle is not stable, and by the fact that signals in traveling to the receiver

often deviate by a few degrees from the great-circle route. Experience shows that it is permissible to employ more horizontal than vertical directivity, but in both cases there is a limit beyond which it is undesirable to concentrate the radiation.

Character of Received Signals.—The short-wave signals received over a relatively long distance usually represent the vector sum of different waves that have traveled over paths of different lengths.[1] As the relative lengths of these paths vary continuously with changing conditions in the ionosphere, the amplitude of the resultant signal varies continuously, causing fading that is frequently very severe. Inasmuch as the difference in path length depends upon the frequency of transmission, the different frequencies contained in a modulated wave will not combine in the same phase, thus giving rise to selective fading that is often very severe. Investigations of selective fading on transatlantic short-wave telephone circuits have shown that at times frequencies differing by as little as 100 cycles fade independently. Relative signal strengths for different modulation frequencies at successive moments are shown in Fig. 356*a* for a typical case of transatlantic short-wave transmission and indicate the highly variable character of the fading.[2]

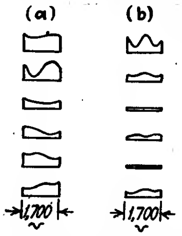


FIG. 356*a*.—Typical fading patterns observed in short-wave transatlantic communication. The different figures in each column show transmission conditions at successive moments over a 1700-cycle band of modulation frequencies.

When the same signal is received on antennas spaced 10 wave lengths or more apart, the fading is ordinarily substantially independent. This fact is taken advantage of in the diversity receiving systems described in Sec. 114 in order to minimize fading.

Short-wave signals ordinarily are found to have both vertically and horizontally polarized components which bear no apparent relation to the polarization of the transmitting antenna, which fade independently of each other, and which ordinarily are not in phase. This situation is the result of the earth's magnetic field, which causes waves refracted by the ionosphere to have their plane of polarization rotated, and which

[1] Thus experiments made with transatlantic signals have indicated that there are ordinarily two to four transmission paths contributing appreciable energy to the receiver. These paths are characterized at the receiving point by different vertical angles commonly ranging from 10 to 25 deg., with occasional paths delivering appreciable energy at angles up to 40 deg. See H. T. Friis, C. B. Feldman, and W. M. Sharpless, The Determination of the Direction of Arrival of Short Radio Waves, *Proc. I.R.E.*, vol. 22, p. 47, January, 1934.

[2] For further data on short-wave fading and quality distortion, see R. K. Potter, Transmission Characteristics of a Short-wave Telephone Circuit, *Proc. I.R.E.*, vol. 18, p. 581, April, 1930.

also splits the wave into several components that travel along different paths with different velocities.

Echo and Multiple Signals.—The fact that short waves can travel to a receiver along paths of different lengths gives rise to echo signals of which a number of kinds have been observed. Thus, if a short-wave train lasting perhaps 10^{-4} sec. is sent out from a transmitter and recorded at the receiving point on an oscillogram, the results will often be as in Fig. 357, in which the same impulse is shown as being received a number of times. When the receiver is within the range of the ground wave, the first received impulse travels along the ground, while the second impulse has reached the receiver by way of the Kennelly-Heaviside layer and so arrives several thousandths of a second later as a result of its longer path. The remaining echoes are apparently the result of multiple-step paths between the earth and the Kennelly-Heaviside layer, or are the result of magneto-ionic splitting.

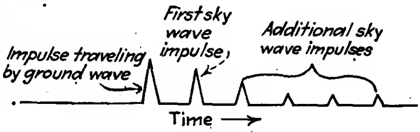


Fig. 357.—Typical oscillogram of received signal when transmitter within ground-wave range sends out a short impulse.

Multiple signals are sometimes observed as a result of waves that have traveled around the earth. The possible paths for such around-the-world multiple signals are shown in Fig. 358, where it is seen that the signals may reach a distant receiving point along either the short or long great-circle path, and that, if the waves traveling in either of these directions make more than one complete circle of the earth, they will be heard again. Multiple round-the-world signals occur regularly on certain short-wave circuits at definite times each year. There are cases on record where the same signal has been heard five times at a distant receiving point, while multiple signals repeated once or twice are relatively common.[1] In order that double signals may reach the receiver by propagation along both long and short great-circle paths, it is necessary that the conditions in the Kennelly-Heaviside layer be roughly the same along both routes. Otherwise a frequency that is satisfactorily propagated along one of the two paths will not be able to reach the receiver along the other path. The same requirement must also be fulfilled in the case

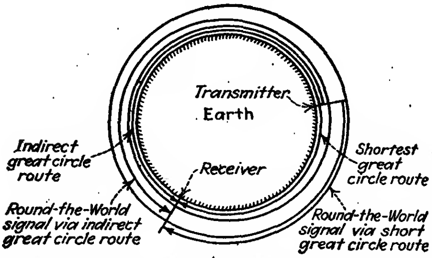


Fig. 358.—Paths followed by multiple and round-the-world signals.

[1] See E. Quäck and H. Mögel, Double and Multiple Signals with Short Waves, *Proc. I.R.E.*, vol. 17, p. 791, May, 1929.

of round-the-world signals. As a result multiple and round-the-world signals are observed only when the part of the great-circle path that lies in darkness is experiencing summer, or when the great circle path coincides very closely with the twilight zone.[1] The time delay of round-the-world signals is in the order of 1/7 sec. and is so great that, when these signals occur, it is necessary to reduce the speed of telegraph transmission to an extremely low value. With radio-telephone signals, the fading and distortion is increased.

Echo signals having a time lag of several seconds have been reported a number of times, and several well-authenticated cases have been observed in which the time lag was several minutes. The cause of such echoes has not been definitely established, but theoretical work by P. O. Pedersen[2] indicates that retardations up to 10 sec. could be accounted for by low group-velocity propagation in the Kennelly-Heaviside layer, while signals with greater retardations can be accounted for only by waves which have traveled great distances in the empty space outside the earth's atmosphere and which by a fortunate combination of circumstances have finally been reflected back to earth by ionized regions either within the influence of the earth's magnetic field or in the vicinity of the sun. These echo signals of long delay, while of extreme theoretical interest, are of little practical importance because they occur so rarely.

124. Propagation of Ultra-high-frequency Waves.[3]—Frequencies too high to be refracted back to earth by the ionosphere, even when leaving the earth along the horizontal, pass on through the ionized region and are lost. The critical frequency above which this occurs depends upon

[1] *Ibid.*; A. H. Taylor and L. C. Young, Studies of High Frequency Radio-wave Propagation, *Proc. I.R.E.*, vol. 16, p. 561, May, 1928.

[2] P. O. Pedersen, Wireless Echoes of Long Delay, *Proc. I.R.E.*, vol. 17, p. 1750, October, 1929.

[3] The literature dealing with the propagation of ultra-high-frequency waves is very extensive. For the reader who desires to pursue the subject further, the following selected bibliography is suggested as a start: Charles R. Burrows, Alfred Decino, and Loyd E. Hunt, Ultra-short-wave Propagation over Land, *Proc. I.R.E.*, vol. 23, p. 1507, December, 1935; Charles R. Burrows and Alfred Decino, Ultra-short-waves in Urban Territory, *Elec. Eng.*, vol. 54, p. 115, January, 1935; L. F. Jones, A Study of the Propagation of Wavelengths between Three and Eight Meters, *Proc. I.R.E.*, vol. 21, p. 349, March, 1933; C. R. Englund, A. B. Crawford, and W. W. Mumford, Some Results of a Study of Ultra-short-wave Transmission Phenomena, *Proc. I.R.E.*, vol. 21, p. 464, March, 1933; C. R. Englund, A. B. Crawford, and W. W. Mumford, Further Results of a Study of Ultra-short-wave Transmission Phenomena, *Bell System Tech. Jour.*, vol. 14, p. 369, July, 1935; J. C. Schelleng, C. R. Burrows, and E. B. Ferrell, Ultra-short-wave Propagation, *Proc. I.R.E.*, vol. 21, p. 427, March, 1933; Bertram Trevor and P. S. Carter, Notes on Propagation of Waves below Ten Meters in Length, *Proc. I.R.E.*, vol. 21, p. 387, March, 1933; J. A. Stratton, The Effect of Rain and Fog on the Propagation of Very Short Radio Waves, *Proc. I.R.E.*, vol. 18,

the maximum electron density in the layer, and so has diurnal and seasonal variations and also changes from year to year. In the daytime the critical frequency is of the order of 30 to 40 mc, while at night it is approximately 15 mc, with a tendency to be higher during the summer and during the active part of the sun-spot cycle. At these high frequencies the ground wave is also very quickly absorbed as a result of earth losses. Satisfactory communication over appreciable distances with very high frequency waves is hence obtained only by utilizing waves passing directly from transmitter to receiver through the space above the ground. However, even when the transmitting and receiving antennas are placed at the highest possible elevation, the range of the signals is limited to moderate distances by the curvature of the earth.

When the distance between transmitter and receiver is small enough for the earth to be considered flat, the situation is as shown in Fig. 359. Here energy may reach the receiver by a direct path between transmitting and receiving antennas or by a route involving reflection from the surface of the ground. The waves traveling through space suffer little attenuation other than that caused by spreading, and therefore have a field strength that is inversely proportional to the distance from the transmitter. At the reflection point the angle of incidence is so near grazing (assuming that the distance of transmission is great compared with the antenna heights) that the reflection is practically complete and takes place with a reversal in phase [see Eq. (216) and Fig. 346]. In spite of this reversal in phase of the reflected wave, the two waves do not cancel each other at the receiver because the direct and indirect paths have different lengths. When the antenna heights are small compared with the distance of transmission, the presence of the indirect reflected wave reduces the resultant field at the receiving point below the free-space value of the direct wave by the factor $4\pi h_s h_r/\lambda d$, where h_s and h_r represent the heights of the sending and receiving antennas above ground, d is the distance between antennas, and λ is the wave length.[1] Since the strength of the direct wave is inversely

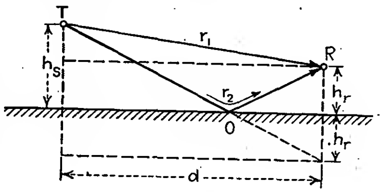


FIG. 359.—Diagram showing the direct and indirect paths by which energy may travel from transmitter to receiver. For the sake of clarity the antenna heights have been greatly exaggerated in comparison with the distance.

p. 1064, June, 1930; R. Jouaust, Some Details Relative to Propagation of Very Short Waves, *Proc. I.R.E.*, vol. 19, p. 479, March, 1931; R. S. Holmes and A. H. Turner, An Urban Field-strength Survey at Thirty and One Hundred Megacycles, *Proc. I.R.E.*, vol. 24, p. 755, May, 1936.

[1] This can be shown as follows: Referring to Fig. 359, it is seen from the dotted construction that: $r_1^2 = (h_s - h_r)^2 + d^2$, and $r_2^2 = (h_s + h_r)^2 + d^2$. For $d \gg$

proportional to distance, the strength of the resultant wave can be expressed by the formula

$$\text{Field at receiver} = \frac{4\pi h_s h_r E_0}{\lambda d^2} \tag{219}$$

where E_0 is the strength of the direct wave at a distance $d = 1$ from the transmitter.[1] It will be noted that, as a consequence of the destructive interference caused by the reflected wave, the received field is inversely proportional to the *square* of the distance and is much less than the free space field. Examination of Eq. (219) shows also that it is highly important for both transmitting and receiving antennas to be as high as possible above the ground, and that, if other things are equal, there is considerable advantage in reducing the wave length. As a consequence, antennas for ultra-high-frequency transmission over considerable distances are preferably located on tall buildings, or even mountain peaks.

Equation (219) assumes that the earth is a perfect reflector and that the direct and indirect paths differ in length by a small fraction of a wave length. Analysis shows that if the height of the lowest antenna exceeds several wave lengths, the error involved in this assumption is small, while with a lower height the imperfect ground reflection causes the received field to be greater than given by Eq. (219).[2]

When ultra-high-frequency waves are propagated over urban territory, it is found that the field strength observed near the surface of the earth is less than that calculated by Eq. (219) by a fixed factor of the order of 0.3 to 0.1. The reason for this appears to be that the field strength existing above the buildings is given by Eq. (219), but that the

---

$(h_s + h_r)$, one can then write

$$r_1 = d + \frac{(h_s - h_r)^2}{2d}, \qquad r_2 = d + \frac{(h_s + h_r)^2}{2d}$$

Consequently the difference in path lengths

$$r_2 - r_1 = \frac{(h_s + h_r)^2 - (h_s - h_r)^2}{2d} = \frac{2h_s h_r}{d}$$

The corresponding phase difference caused by the path difference is $2\pi \cdot 2h_s h_r/\lambda d = 4\pi h_s h_r/\lambda d$ radians. It is because of this angle that the direct and indirect rays fail to cancel, so that the resultant of the two waves is $2 \sin (2\pi h_s h_r/\lambda d)$ times the amplitude of one of the waves. When the angle is so small that the sine of the angle equals the angle in radians, the reduction factor becomes $4\pi h_s h_r/\lambda d$.

[1] When the transmitting antenna radiates a field proportional to the cosine of the angle of elevation, then $E_0 = 300\sqrt{P}$ mv per meter, where E_0 is the direction of maximum radiation, P is the transmitter power in kilowatts, and d is in kilometers.

[2] For further discussion of the effect of imperfect reflection by the earth, see the appendix in the paper by Burrows, Decino, and Hunt, *loc. cit.*

scattering required to produce signals at street level introduces a certain additional loss factor.[1]

Effect of Earth Curvature.—When the distance between transmitter and receiver is considerable, it is necessary to take into account the curvature of the earth. This curvature reduces the received field strength below the values that would be obtained for a flat earth, but experience shows that the correction required is so small as to be negligible as long as a straight-line path exists between transmitter and receiver. If a straight-line path does not exist, it is still possible, however, for energy to reach the receiver either as a result of refraction by the earth's atmosphere or by diffraction of energy around the curved earth, although the received energy is less than if a straight-line path were present.

The refraction by the earth's atmosphere at ultra-high frequencies comes about because the variation of atmospheric pressure, temperature, and moisture content with height causes the refractive index of the atmosphere to decrease with elevation, and this tends to bend the waves back toward the earth according to Eq. (209*b*). The amount of curvature that results varies from time to time with the atmospheric conditions, but on the average it is equivalent to assuming that the earth's radius is increased by 25 to 35 per cent. As a consequence of this refraction, it is possible to obtain direct-ray propagation of energy between transmitter and receiver under conditions where a straight-line path falls slightly below the earth's surface, as in Fig. 360.

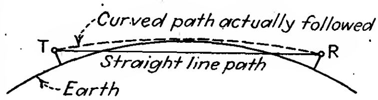


FIG. 360.—Diagram illustrating how the refraction in the earth's atmosphere permits direct ray transmission at ultra-high frequencies even when the straight-line path is intercepted by the earth's curvature.

The range of the direct rays depends upon the heights h_s and h_r of the transmitting and receiving antennas, respectively, and the effective radius of the earth, according to the formula

$$\left.\begin{array}{l}\text{Maximum possible distance for}\\ \text{direct-ray transmission}\end{array}\right\} = 1.225k(\sqrt{h_s} + \sqrt{h_r}) \qquad (220)$$

where the antenna heights are in feet, the distance is in miles, and k represents the ratio of effective to actual earth radius and takes into account the refraction in the earth's atmosphere. The result of Eq. (220) is illustrated in Fig. 361 for $k = 1.33$, which shows that direct-ray transmission over considerable distances is possible only when the antennas are located upon mountain peaks. Even then the range is comparatively modest compared with that obtainable with waves of lower frequency where refraction from the ionosphere can be utilized.

[1] See Burrows and Decino, *loc. cit.*

Theoretical analysis indicates that the earth curvature reduces the received signal below the value calculated by Eq. (219) by the factor given by Fig. 362.[1] This factor takes into account the refraction in the atmosphere and also the diffraction of the energy around the curved surface. Under practical conditions the reduction factor of Fig. 362 is negligible as long as a straight-line path exists, but at greater distances it decreases rapidly and the signals soon become unusable because of fading, as mentioned below.

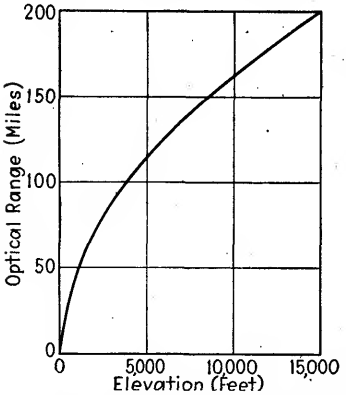


FIG. 361.—Maximum possible optical range between an elevated point and the surface of the earth, assuming that atmospheric refraction increases the effective value of the earth's radius by a factor of 1.33. The maximum possible optical range between two elevated points is the sum of the ranges as obtained in the above diagram for the two heights involved.

Miscellaneous Considerations.—The transmission paths at ultra-high frequency are extremely stable as long as a straight-line path is possible between transmitting and receiving points. There is then no fading, and the polarization of the wave at the transmitter is maintained very accurately. However, when the transmission distance is so great that it is necessary to depend upon the refraction in the earth's atmosphere to obtain direct-ray transmission to the receiver, fading frequently occurs as a result of variations in

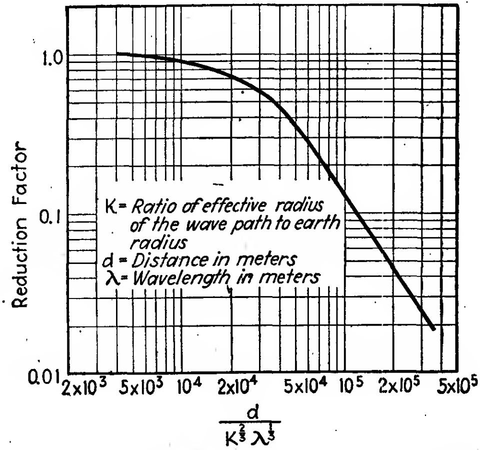


FIG. 362.—Reduction factor caused by diffraction of wave around the surface of the earth.

the earth's atmosphere which change the amount of refraction that the wave suffers. This fading becomes very severe when the transmission

[1] See Burrows, Decino, and Hunt, *loc. cit.*

distance approaches or exceeds the maximum possible distance of direct-ray transmission.

The discussion of ultra-high-frequency transmission given above has made no distinction between horizontally and vertically polarized waves. This is because, at least to a first approximation, the polarization has only a secondary effect.

Inasmuch as the only radiated energy that has any chance of reaching the receiver is that sent out directly toward the receiver, it is possible to increase greatly the strength of the received signal by employing a transmitting antenna that directs the energy correspondingly. The amount of directivity that can be employed to advantage is extremely great, since the waves follow a stable, unvarying path.

125. Use of Radio Waves in Investigations of the Upper Atmosphere. Since the way in which radio signals propagate is dependent on the conditions in the ionosphere, it is possible by working backward from observed propagation characteristics of radio signals to obtain information regarding the probable nature of the ionized layer, and hence indirectly to obtain data on the composition of the upper atmosphere.

Pulse Experiments.—One method commonly used in making such investigations consists in transmitting a short wave train lasting about 10^{-4} sec. and taking a record on an oscillogram of the signal as received at a point within range of the ground wave. Since the wave that reaches the receiver after refraction by the Kennelly-Heaviside layer must travel an appreciably longer path to reach the receiver than does the ground wave, there will be a time interval between the pulses arriving over the two routes. The length of this time interval is a measurement of the difference in path lengths and can be used to estimate the height of the layer.[1] The received records commonly show a number of returned pulses with different time delays as indicated in Fig. 357. These may be the result of simultaneous reflection or refraction from several layers, or the result of double refraction produced as a consequence of the earth's magnetic field, or they may be due to multiple reflections such as two round trips between the earth and the ionosphere.

The most effective means of utilizing the pulse method in investigating the ionosphere is to vary the transmitter frequency, either continuously or in steps, and to obtain the variation of apparent or virtual height as a function of frequency.[2] An example of a simple record of this type is

[1] This method was first proposed by G. Breit and M. Tuve, A Test of the Existence of the Conducting Layer, *Phys. Rev.*, vol. 28, p. 554, September, 1926.

[2] For further information on the technique of making tests, see T. R. Gilliland, Note on a Multifrequency Automatic Recorder of Ionosphere Heights, *Proc. I.R.E.*, vol. 22, p. 236, February, 1934; Lal C. Verman, S. T. Char, and Aijaz Mohammed, Continuous Recording of Retardation and Intensity of Echoes from the Ionosphere, *Proc. I.R.E.*, vol. 22, p. 906, July, 1934.

illustrated in Fig. 363. This shows that at a frequency of about 2000 kc reflections were returned from a layer having a height of 110 km (the E layer), but as the frequency was increased the apparent-layer height first increased gradually, and then suddenly jumped to about 300 km at 3000 kc. This jump takes place at the critical frequency for the E layer, *i.e.*, the frequency at which the wave just barely reaches the point in the layer where the electron density is maximum. For vertical incidence the corresponding refractive index at the point of maximum electron density is zero, so from Eq. (210) the maximum electron density in the E layer for this case is $(3000)^2/81 = 1.1 \times 10^5$ electrons per cubic centimeter. As the frequency was increased still further, the virtual height first decreased to about 190 km, and then rose to a second peak at 4000 kc corresponding to the critical frequency at which the waves were just able to penetrate to the maximum electron density in the F_1 layer.

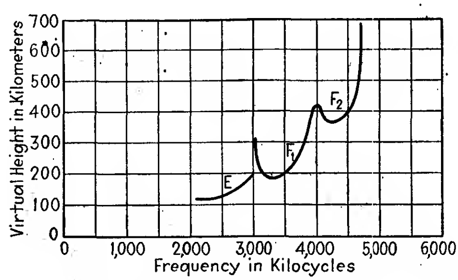


FIG. 363.—Typical curves of virtual height as a function of frequency, as obtained in ionospheric investigations.

The corresponding maximum electron density was 2.0×10^5 electrons per cubic centimeter. As the frequency was increased still more, another critical frequency was obtained at 4700 kc corresponding to a maximum density of 2.7×10^5 electrons per cubic centimeter in the F_2 layer. At frequencies above this value waves striking the layer at vertical incidence penetrated through the layer and did not return to earth.

The heights obtained from the delay of the sky-wave pulses are the "virtual" or apparent heights, and are always greater than the maximum height reached by the wave because of the fact that in the ionized region the waves travel more slowly than in free space. The actual group velocity is equal to the velocity of light multiplied by the refractive index. The slowing down is hence very great near the critical frequency, for here the refractive index approaches zero and the velocity becomes very low. This accounts for the apparent rise in height near the critical frequency which appears in Fig. 363. The actual heights of the points of maximum electron densities for this case are of the order of 100, 190, and 375 km for the E, F_1, and F_2 layers, respectively.

A number of other methods of investigating the ionosphere have been devised and are used under certain circumstances. Thus Appleton and Barnett[1] investigated the ionized layer by varying the transmitted frequency and observing the variations that occurred in the signal strength at a point within range of the ground wave. Because of the different path lengths of sky and ground waves, the relative phase of the two component waves reaching the receiver will alternate between the same phase and phase opposition as the frequency is varied, and the increment in frequency that is required to change the relative phases by 180° can be used to estimate the layer height. Theoretical analysis indicates that the layer height obtained in this way is the virtual height given by the pulse method of Breit and Tuve. This method of determining the height of the ionosphere is particularly satisfactory at the lower frequencies, where the pulse method is not practicable, but is not satisfactory at the higher frequencies because with several refracted waves (as is common at the higher frequencies) the observed results are too complex to permit of interpretation.

Appleton and his co-workers have also made extensive studies of other characteristics of the received waves. With receivers located several hundred kilometers from the transmitter, they have deduced the virtual height of the ionized layer by measuring the angle of incidence at which the down-coming waves strike the earth. They have also worked out the technique by which it is possible to deduce the characteristics of an elliptically polarized wave, including both the magnitude and relative phases of the horizontal and vertical components and also the sense of rotation.

Still another way of determining the height of the Kennelly-Heaviside layer consists in observing the variations in field intensity as the distance between transmitter and receiver is varied. Thus Hollingworth[2] has found that at moderate distances from long-wave transmitters the signal strength alternately decreases and increases as the distance between transmitter and receiver is varied. These variations in signal strength result from alternate reinforcement and cancellation between sky and ground waves.

126. Relation of Solar Activity and Meteorological Conditions to the Propagation of Radio Waves.—The fact that the propagation of all except the very shortest radio waves depends to a marked extent upon the conditions in the ionosphere would lead one to expect some relation

[1] See E. V. Appleton, Some Notes on Wireless Methods of Investigating the Electrical Structure of the Upper Atmosphere, *Proc. Phys. Soc.*, vol. 41, Part II, p. 43, December, 1928.

[2] See J. Hollingworth, Propagation of Radio Waves, *Jour. I.E.E. (London)*, vol. 64, p. 579, May, 1926.

to exist between solar activity and meteorological conditions on the one hand, and wave propagation on the other, and this is the case.

The most striking relation of this type is the abnormal propagation characteristics of radio waves that always accompany magnetic storms. A magnetic storm is characterized by a rapid and excessive fluctuation of the earth's magnetic field which begins almost simultaneously over the entire earth with full intensity and then gradually subsides in 3 or 4 days.[1] Magnetic storms occur more or less irregularly, although showing a tendency to reoccur at intervals of 27.3 days, which is the period of rotation of the sun.

During a magnetic storm the daytime field strength on long waves is increased above normal, the sunset drop in signal intensity disappears, and the night field is subnormal. These effects are illustrated in Fig. 364 and are more pronounced as the frequency of transmission is increased.

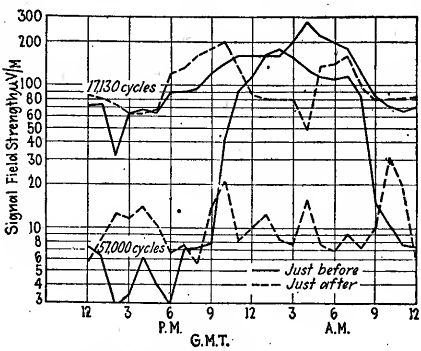


FIG. 364.—Effect of magnetic storm on low-frequency waves, showing how the day field strength is increased and the night field strength reduced by the magnetic storm.

The propagation of high-frequency radio waves is very adversely affected by magnetic storms, particularly when the transmission path passes near the polar regions. Thus a severe magnetic storm ordinarily makes the short-wave circuits across the north Atlantic completely inoperative for a period of several days and causes subnormal field strengths for a week or more. The adverse effect is much less, however, when the entire transmission path is nowhere near the poles. This is strikingly brought out by Fig. 365, which shows the signal strength over the New York-London and the New York-Buenos Aires short-wave circuits during and after the same magnetic storm.

In addition to the abnormalities of wave propagation which are associated with magnetic storms, there also appears to be some relation between sun spots and radio-wave propagation. Thus yearly averages of field strength of long-wave signals arriving from distant transmitters

[1] An excellent summary of the principal solar phenomena that are of importance in wave propagation is given by Clifford N. Anderson, Correlation of Long-wave Trans-atlantic Radio Transmission with Other Factors Affected by Solar Activity, *Proc. I.R.E.*, vol. 16, p. 297, March, 1928. Also see Austin Bailey and H. M. Thomson, Transatlantic Long-wave Telephone Transmission and Related Phenomena from 1923 to 1935, *Bell System Tech. Jour.* vol. 14, p. 680, October, 1935; A. M. Skellett, On the Correlation of Radio Transmission with Solar Phenomena, *Proc. I.R.E.*, vol. 23, p. 1361, November, 1935.

correlate surprisingly well with yearly averages of sun-spot numbers, as is brought out by Fig. 366.[1] In addition to the eleven-year sun-spot cycle apparent in Fig. 366, there is also a 27.3-day sun-spot cycle corresponding to the period of rotation of the sun. While day-to-day signal strength appears to be independent of day-to-day variations in sun-spot activity, Pickard has found that, when the signal strength over a number of successive 27.3-day cycles is averaged, a very decided relation exists, as is illustrated in Fig. 367.[2]

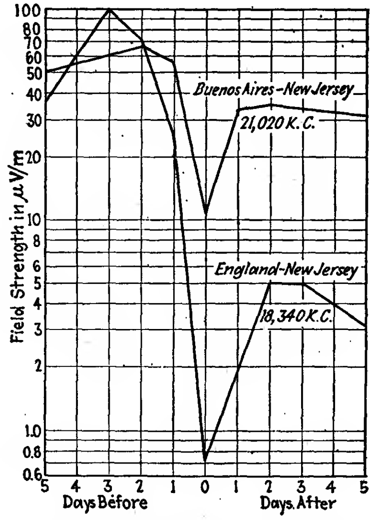


FIG. 365.—Effect of magnetic storm on the propagation of short waves. The drop in field strength is much greater on the east-west than on the north-south circuit.

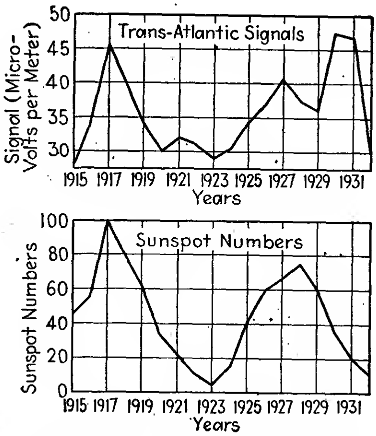


FIG. 366.—Diagram showing the close correlation between sun-spot numbers and yearly average of long-wave signal strength.

The variations in wave propagation which take place with sun spots, magnetic storms, etc., are the result of corresponding variations in the conditions existing in the ionosphere. Experimental data concerning the exact nature of these variations in the ionosphere are very fragmentary, although certain trends appear to exist. Thus the highest frequency that is returned to earth at great distances during the day varies more or less in synchronism with the eleven-year sun-spot cycle, being highest when the sun spots are most numerous. Likewise the electron density in the F_1 layer apparently decreases on magnetically disturbed days, and

[1] See E. B. Judson, Low-frequency Radio Receiving Measurements at the Bureau of Standards, *Proc. I.R.E.*, vol. 21, p. 1354, September, 1933.

[2] See Greenleaf W. Pickard, Correlation of Radio Reception with Solar Activity and Terrestrial Magnetism, Part II, *Proc. I.R.E.*, vol. 15, p. 749, September, 1927.

the absorption that the sky waves suffer tends to be greater during magnetic storms and on magnetically disturbed days than otherwise.

The effect of solar eclipses on the ionosphere can be investigated with the aid of pulse signals. Results indicate that the changes caused by the eclipse are in phase with the optical eclipse and that during the period of totality the maximum electron density in the ionosphere decreases appreciably.[1]

Some relation appears to exist between the strength of received signal, atmospheric temperature, barometric pressure, and weather conditions, although the correlation is not high and tends to be obscured by other influences.[2]

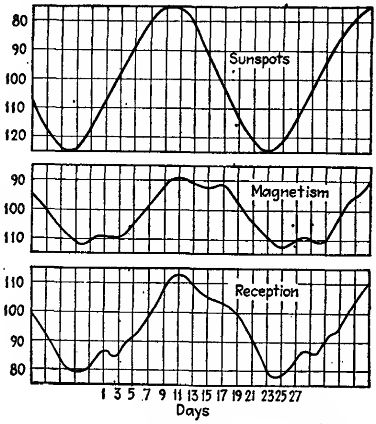


FIG. 367.—Average of sun-spot numbers, magnetic character of days, and radio reception on 1330 kc for eight solar rotations. These curves have been smoothed by the use of a 13-day moving mean.

127. Noise and Static.—The output of a sensitive radio receiver nearly always contains a background of rumbles, crashes, rattles, etc., which disappears when the antenna is disconnected. This noise is the result of voltages induced in the antenna by either natural or man-made sources of interference and is often of sufficient magnitude to be the practical factor determining the minimum usable signal.

The principal sources of man-made noise are high-voltage power lines, ignition systems of airplane and automobile motors, brush motors, and electrical appliances. The noise from these sources is distributed throughout the entire frequency range used in radio communication and is always strong in cities and particularly in industrialized areas. There is very little that can be done to minimize the general level of man-made noise other than to suppress unusually bad localized sources of interference. Satisfactory reception in populated areas is hence obtained only from strong signals.

[1] A summary of eclipse effects is given by E. V. Appleton and S. Chapman, Report on Ionization Changes during a Solar Eclipse, *Proc. I.R.E.*, vol. 23, p. 658, June, 1935.

[2] Thus see Greenleaf W. Pickard, Some Correlations of Radio Reception with Atmospheric Temperature and Pressure, *Proc. I.R.E.*, vol. 16, p. 765, June, 1928; L. W. Austin and I. J. Wymore, Radio Signal Strength and Temperature, *Proc. I.R.E.*, vol. 14, p. 781, December, 1926; R. C. Colwell, Weather Forecasting by Signal Radio Intensity, *Proc. I.R.E.*, vol. 18, p. 533, March, 1930; R. C. Colwell, Cyclones, Anticyclones, and the Kennelly-Heaviside Layer, *Proc. I.R.E.*, vol. 21, p. 721, May, 1933.

Radio waves generated by natural causes are referred to as static and produce the familiar clicks, rumblings, crashes, etc., sometimes heard in all radio receivers. Static normally has its origin in thunderstorms and similar natural electrical disturbances and is in the form of impulses, the energy of which is distributed throughout the range of useful radio frequencies.[1] The energy level of static decreases as the frequency increases, ordinarily being very great at the lower radio frequencies and so small as to be unimportant at ultra-high frequencies. Since static is fundamentally a radio signal containing frequency components distributed over a wide range of frequencies, the static within any frequency range is propagated over the earth in the same way as ordinary radio signals of the same frequency. Thus static impulses travel great distances under favorable conditions, arrive at a receiving point from a definite direction, and possess diurnal and seasonal variations in intensity as a result of corresponding variations in wave propagation.

It has been found that some of the static observed in the northern hemisphere is produced by local thunderstorms, but that a surprisingly large part of static interference originates in the tropics. Thus directional observations on long-wave static in Maine give a general southwesterly origin pointing toward the Gulf of Mexico and Texas, while similar observations in Europe indicate sources in Africa.[2] It has also been found that land areas, particularly mountains, are usually the most important sources of static, and that the static is worst in the summer season.

Low and Moderately Low Frequency Static.—At low radio frequencies the static intensity is high because most of the energy of a static impulse is concentrated on the lower radio frequencies and because at low frequencies radio waves propagate great distances with small attenuation.[3]

[1] Information on the nature of the static pulses produced by thunderstorms is given by H. Norinder, Cathode-ray Oscillographic Investigations on Atmospherics, *Proc. I.R.E.*, vol. 24, p. 287, February, 1936.

At short waves there is a relatively weak hisslike static having an interstellar origin in the direction of the Milky Way. See Karl G. Jansky, A Note on the Source of Interstellar Interference, *Proc. I.R.E.*, vol. 23, p. 1158, October, 1935.

[2] The connection between storm areas and static is strikingly brought out by observations made in Maine by engineers of the American Telephone and Telegraph Company, in which it was found that storms within several thousand miles could be readily followed for days by making directional observations on static. See A. E. Harper, Some Measurements on the Directional Distribution of Static, *Proc. I.R.E.*, vol. 17, p. 1214, July, 1929; S. W. Dean, Correlation of Directional Observations of Atmospherics with Weather Phenomena, *Proc. I.R.E.*, vol. 17, p. 1185, July, 1929.

[3] Thus the same static impulse has been observed at Berlin, Germany, and the Hawaiian Islands. See M. Bäumler, Simultaneous Atmospheric Disturbances in Radio Telegraphy, *Proc. I.R.E.*, vol. 14, p. 765, December, 1926; also see S. W. Dean, Long-distance Transmission of Static Impulses, *Proc. I.R.E.*, vol.19, p. 1660, September, 1931.

The intensity of long-wave static becomes greater as the frequency is reduced, and in northern latitudes is greater at night and summer than in the daytime and winter, respectively. The curves of Fig. 368 summarize the more outstanding features of long-wave static as observed in the northern hemisphere.[1]

At moderate frequencies, such as those in the broadcast range, a large fraction of the static observed during the day is of local origin because of the poor propagation of such frequencies during daylight hours. At night, however, the lower attenuation causes static impulses of distant origin to be heard, with the result that the noise level is ordinarily greater at night than in the daytime.

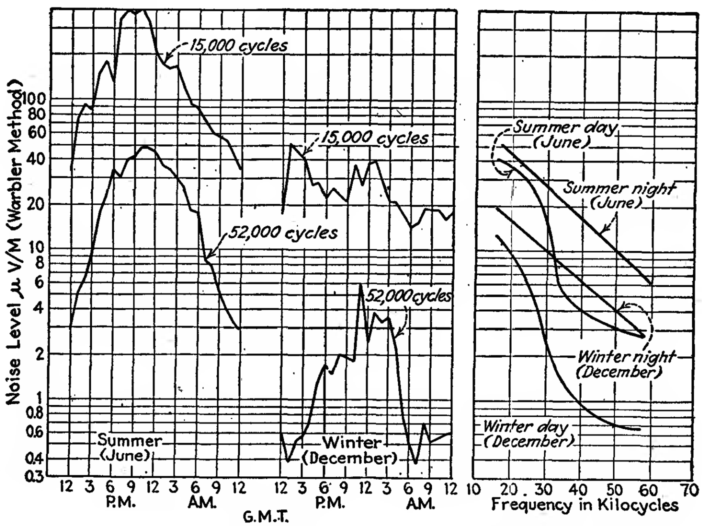


FIG. 368.—Curves summarizing behavior of low-frequency static received in Maine.

Short-wave Static.—The static intensity at short waves, *i.e.*, frequencies from 6000 to 30,000 kc, is much less than at lower frequencies, and during a good part of the time is of the same order of magnitude as the noise level of a typical radio receiver. Investigations of high-frequency static show evidences of localized sources similar to those observed at lower frequencies, and it appears that a large fraction of high-frequency static represents waves that have traveled considerable distances. As a result, high-frequency static at any given wave length shows diurnal and seasonal

[1] Excellent discussions of long-wave static are given by L. W. Austin, The Present Status of Radio Atmospheric Disturbances, *Proc. I.R.E.*, vol. 14, p. 133, February, 1926; Lloyd Espenschied, C. N. Anderson, and Austin Bailey, Trans-atlantic Radio Telephone Transmission, *Proc. I.R.E.*, vol. 14, p. 7, February 1926; and R. A. Watson Watt, Present State of Knowledge of Atmospherics, abstract in *Exp. Wireless and Wireless Eng.*, vol. 5, p. 629, November, 1928.

variations in intensity which are identical with the corresponding variations in the strength of long-distance short-wave signals of the same frequency. This is well illustrated by the curves of Fig. 369, which show that during the day the static is strongest on the frequency best suited for daylight transmission (a high frequency), while at night a reversal of the situation takes place because a low frequency is best for short-wave night transmission.[1]

High-frequency static is closely correlated with low-frequency static. Investigations show that both arrive from the same directions and that sometimes it is even possible to correlate individual high- and low-frequency impulses.

Very little static is found on frequencies too high to be refracted from the ionosphere. At these frequencies the range of the signals is so limited that all static must be of relatively local origin, and even then there is very little static since it appears that nature, like man, finds difficulty in generating waves at such high frequencies.

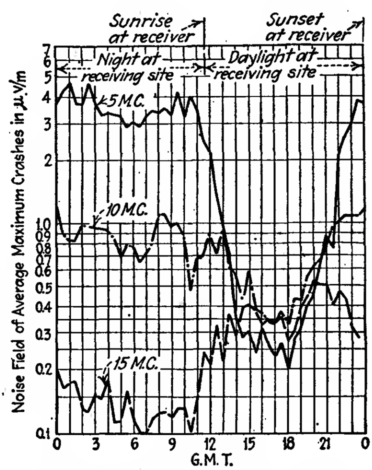


FIG. 369.—Diurnal variation of short-wave static received at Doria, Fla., in December. Note how the noise level on each frequency is highest at the part of the day most favorable for long-distance propagation of waves of the same frequency.

Means of Overcoming Static.—The only successful means that have been found for minimizing static interference are to make the frequency band to which the receiver responds as narrow as possible and to use directional receiving antennas. Since static energy is more or less uniformly distributed through the frequency spectrum, it is obvious that the amount of static energy picked up by a receiver is almost directly proportional to the frequency range to which the receiver responds. This range should therefore be no wider than is necessary to accommodate the side bands of the desired signal.[2]

Directional receiving systems are of advantage in eliminating static interference when the interference and the desired signal arrive from

[1] The result of a very thorough investigation of high-frequency noise is reported by R. K. Potter, High-frequency Atmospheric Noise, *Proc. I.R.E.*, vol. 19, p. 1731, October, 1931. Also see R. K. Potter, Estimate of the Frequency Distribution of Atmospheric Noise, *Proc. I.R.E.*, vol. 20, p. 1512, September, 1932.

[2] See John R. Carson, Selective Circuits and Static Interference, *Trans. A.I.E.E.*, vol. 43, p. 789, 1924. This is a classical paper in which it is shown that, if static is a random series of impulses, then the amount of static energy absorbed by a receiver is directly proportional to the frequency range to which the receiver responds.

different directions. It has been found that in east-west transmission at high latitudes very great gains in signal-to-noise ratio are nearly always realized by using directive receiving antennas because the major sources of static heard on almost any frequency are to the south and so are not in the direction of the transmitting station. After the full benefits of narrow-band reception and directional receiving antennas have been realized, the only remaining possibility for improvement in the signal-to-static ratio is to increase the transmitter power in order to override the interference, or to move the receiver to a location where the static intensity is less.

Much effort has been expended in attempting to devise "static eliminators," but all of these devices have been failures. The reason is that static is a radio wave similar in all respects to the signal that is to be received, and any balancing scheme that balances out static will also balance out the received signal. In cases where apparent improvement has been obtained it can be shown that this is the result of improved selectivity (*i.e.*, a narrowing of the width of the response band of the receiver) rather than because of any balancing action that is present. In telegraph receivers the width of the response band of ordinary receivers is much greater than the minimum necessary to accommodate the side-bands represented by the dots and dashes, so that considerable improvement in signal-to-static ratio of code signals can ordinarily be obtained by increased receiver selectivity.[1]

The use of frequency modulation at ultra-high frequencies has recently been proposed as a means of reducing static interference. Preliminary results have been quite promising, and theoretical analysis indicates that frequency modulation is superior to amplitude modulation in suppressing noise, provided the ratio of signal to noise voltages exceeds a certain limiting value.[2]

128. Reciprocal Relations in Wave Propagation.—From the Rayleigh-Carson reciprocity theorem discussed in Sec. 130 it follows that the transmitter and receiver can under ordinary conditions be interchanged without producing any effect. This is equivalent to stating, among other things, that reversing the direction of transmission between two fixed points does not affect the propagation, and that the angle at which the waves arrive at the receiving antenna is the same as the vertical angle at

[1] See John R. Carson, The Reduction of Atmospheric Disturbances, *Proc. I.R.E.*, vol. 16, p. 966, July, 1928. In this paper Carson shows the fallacy behind a number of proposed methods of balancing out static interference while still preserving the signal.

[2] See E. H. Armstrong, A Method of Reducing Disturbances in Radio Signalling by a System of Frequency Modulation, *Proc. I.R.E.*, vol. 24, p. 689, May, 1936; Murray G. Crosby, "Frequency Modulation Noise Characteristics," *Proc. I.R.E.*, vol. 25, p. 472, April, 1937.

which the same waves left the transmitting antenna. These reciprocal relations are exact except for propagation through an ionized medium in the presence of a magnetic field, and so apply to ground-wave and direct-ray transmission without any restrictions. However, even when the ionosphere plays an important part in the wave transmission, the reciprocal relations can still be expected to hold when they are applied to values averaged over a short period of time rather than to instantaneous conditions.

Problems

1. A police radio transmitter is to be designed to provide coverage for a small city and surrounding countryside. Assuming that 80 per cent of the power delivered to the transmitting antenna is actually radiated, that the radiated field is proportional to the cosine of the angle of elevation, and that $\sigma = 0.5 \times 10^{-13}$ and $\epsilon = 20$, determine the transmitter power required to make the ground wave have a strength of 100 μv per meter at a distance of 16 km when the frequency is: (*a*) 1690 kc, (*b*) 2500 kc.

2. A broadcast station operating at a frequency of 1000 kc delivers 50 kw to an antenna which radiates 80 per cent of this power and which has a directional characteristic such that the field radiated along the horizontal is 1.28 times as great as in antennas in which the field is proportional to the cosine of the angle of elevation. Assuming that the earth conductivity is 10^{-13} e.m.u. and that the dielectric constant is 15, calculate the distances from the transmitter at which the strength of the ground wave will be 1 mv per meter, 0.5 mv per meter, 0.2 mv per meter, 0.1 mv per meter, and 0.05 mv per meter.

3. A series of field-strength measurements about a broadcast station operating at 900 kc shows that at a distance of 20 miles the strength of the ground wave is 0.25 of the value calculated on the basis of zero ground losses. Deduce the earth conductivity for this case, assuming a reasonable value for the dielectric constant.

4. On the assumption that the ionosphere is the result of radiations from the sun which ionize the earth's atmosphere, explain why it would be unreasonable to expect the electron densities at several thousand miles above the earth, or at the earth's surface, to be as great as at some intermediate height.

5. The earth's magnetic field has no effect upon a wave passing through the ionosphere provided the wave is so oriented that the electrostatic lines of force of the wave are parallel with the magnetic flux lines of the earth. Explain.

6. Calculate and plot the refractive index of the F_1 layer as a function of frequency up to the frequency for which $\mu = 0$, assuming that the electron density is 2.2×10^5 electrons per cubic centimeter (corresponding to summer noon at Washington as in Fig. 337).

7. From the data given in Fig. 337 calculate and plot as a function of time of day the highest frequency that on the average could be used for short-distance communication (corresponding to a sky wave striking the ionosphere with nearly vertical incidence) at Washington, D. C., in June, 1933, during the hours 8 A.M. to 6 P.M.

8. Calculate the highest frequency that will be returned to earth under any conditions where the ionosphere conditions correspond to those at Washington in June (see Figs. 336 and 337), when the time of day is (*a*) noon, (*b*) midnight.

9. Calculate the magnitude and phase angle of the reflection coefficient for horizontally polarized waves in soil for which the dielectric constant is 20 and the conductivity is 10^{-13} e.m.u. Use frequencies of 100, 1000, and 10,000 kc, and assume vertical incidence ($\delta = 90°$).

10. At broadcast frequencies (550 to 1500 kc) the sky-wave attenuation is quite small at night and very high in the daytime. Describe ionospheric conditions which could explain this behavior and which would at the same time not be inconsistent with Figs. 336 and 337.

11. From the characteristics of the ionosphere explain: (*a*) how it comes about that low-frequency waves suffer less absorption in the ionosphere in the day than waves of broadcast frequencies, and (*b*) why there is less diurnal and seasonal variation in Fig. 347 for 17,300 cycles than for 54,500 cycles.

12. Determine the transmitter power (assuming that the radiated field is proportional to the cosine of the angle of elevation) required to produce a ground-wave field strength of 1 mv per meter at a distance of 200 km when the soil conductivity is 10^{-13} e.m.u. and the dielectric constant is 15, for frequencies of 550, 1000, and 1500 kc. From the results discuss the economic feasibility of obtaining large ground-wave coverage at the different broadcast frequencies.

13. Determine the transmitter power (assuming that the radiated field is proportional to the cosine of the angle of elevation) required to produce a ground-wave field strength of 1 mv per meter at 200 km when the frequency is 550 kc and the earth conductivity is 10^{-11} (sea water), 10^{-13} (good soil), and 10^{-14} (poor soil).

14. Draw curves corresponding to Fig. 351*c*, but giving ideal directional characteristics for different transmitter powers with constant soil conditions.

15. A proposed broadcast transmitter is to deliver 50 kw of power to an antenna that has a directional characteristic as shown in Fig. *A*. The field strength along the ground at a distance of 1 mile (which is so close that ground losses can be neglected) is expected to be 1750 mv per meter. The transmitter frequency is to be 1000 kc.

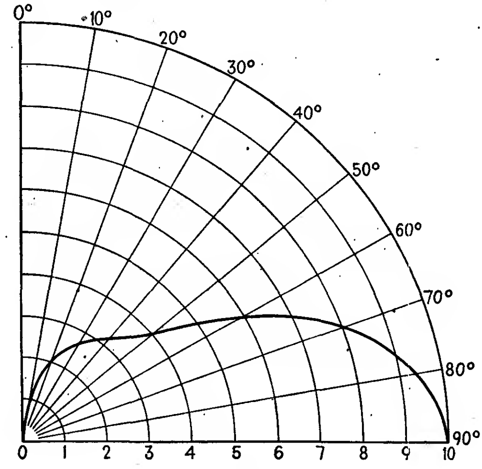


FIG. *A*.

a. Plot a curve of ground-wave field strength as a function of distance up to 750 km for earth conductivities of 1×10^{-13} and 0.4×10^{-13}.

b. Calculate sky-wave field strength as a function of distance up to 750 km by making reasonable assumptions as to the ionospheric action, and plot the results upon the same curve sheet as used for (*a*).

c. Discuss the resulting coverage for both day and night, including consideration of such factors as day and night primary-service areas, location of high-distortion area, amount of rural night coverage obtained, etc.

16. Calculate and plot skip distance as a function of frequency for ionospheric conditions corresponding to those given in Figs. 336 and 337 for noon and midnight in December.

17. Suggest short-wave frequencies suitable for communication at noon and at midnight, over distances of 500 km and 5000 km, assuming that the ionospheric conditions are those shown in Figs. 336 and 337 for June.

18. Show that the data given in Figs. 336 and 337 are consistent with the fact that the optimum frequency for long-distance short-wave communication is much lower at night than in the day, but differs only slightly in winter from summer.

19. Short-wave communication is to be carried on at noon between two points 200 km apart. Determine the approximate vertical angle at which the transmitted energy should be concentrated, assuming that the ionospheric conditions correspond to those for December in Figs. 336 and 337 and that the transmitter frequency is (*a*) 2,750 kc, (*b*) 4,000 kc.

20. When the transmitting and receiving antennas are close together so that the reflected wave reaching the receiver strikes the ground with an appreciable angle of incidence, Eq. (219) no longer holds, but rather the received field strength alternately decreases and increases as the height of either antenna is increased. Explain.

21. The antenna for a proposed television transmitter is to be located at the top of a building 750 ft. high. (*a*) Over what distance is direct-ray coverage possible, assuming that the average receiving antenna will have a height of less than 30 ft.? (*b*) Over what distance is a straight-line path possible, assuming the same receiving antenna?

22. An ultra-high-frequency transmitter operating at a wave length of 5 meters with a power of 100 watts is to be used for communication between two points 50 miles apart. The height of the transmitter antenna is 400 ft. (*a*) Determine the minimum height of the receiving antenna for which direct-ray reception is possible. (*b*) Estimate the field strength received by this antenna, assuming the transmitting antenna radiates a field proportional to the cosine of the angle of elevation and taking into account the earth curvature with the aid of Fig. 362.

23. What transmitter power is required to deliver a 0.050 mv per meter signal at 5 meters when the transmitting and receiving antennas are both 50 ft. high, the distance is 10 miles, and the radiated field is proportional to the cosine of the angle of elevation?

24. When pulse signals are being returned from the F_1 layer, the virtual height is greater when the frequency is just barely high enough to permit penetration of the E layer than when a higher frequency is used, even though the waves actually penetrate farther up into the F_1 layer the higher the frequency. Explain.

25. Explain why single side-band transmission of radio-telephone signals has been found to be an effective means of reducing the effects of static at both high and low frequencies.

CHAPTER XV

ANTENNAS

129. Fundamental Laws of Radiation.—An understanding of the mechanism by which energy is radiated from a circuit and the derivation of equations for expressing this radiation quantitatively involve conceptions that are unfamiliar to the ordinary engineer.[1] The mathematical formulas that express the results of such an analysis are, however, quite simple and understandable. Thus the strength of the field radiated from an elementary length of wire δl carrying a current I (see Fig. 370) is given by the formula:

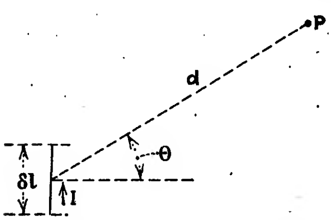


FIG. 370.—Elementary doublet consisting of a length of wire δl carrying a current I.

$$\epsilon = \frac{60\pi}{d\lambda}(\delta l)I \cos \omega\left(t - \frac{d}{c}\right) \cos \theta$$

$$= \frac{60\pi}{dc}f(\delta l)I \cos \omega\left(t - \frac{d}{c}\right) \cos \theta \quad (221)$$

where

ϵ = the strength of the wave in volts per meter

δl = the length of wire from which the radiation takes place, measured in the same units as λ

$I \cos (wt + 90°)$ = current flowing in the wire in amperes

d = distance from P to the antenna in meters

θ = angle of elevation of point at which field is desired with respect to a plane perpendicular to the conductor δl

f = frequency of current

$\omega = 2\pi f$

t = time

c = velocity of light $= 3 \times 10^8$ meters per second

λ = wave length corresponding to frequency f.

[1] For an elementary mathematical analysis of the radiation phenomenon, see R. R. Ramsey and Robert Dreisback, Radiation and Induction, *Proc. I.R.E.*, vol. 16, p. 1118, August, 1928; R. R. Ramsey, Radiation and Induction, *Proc. I.R.E.*, vol. 21, p. 1586, November, 1933. A more advanced treatment is given by G. W. Pierce, "Electric Oscillations and Electric Waves," McGraw-Hill Book Company, Inc., New York, 1920.

The radiated field ϵ varies directly as the current I, the frequency f, the doublet length δl, and the cosine of the angle of elevation, and is inversely proportional to the distance d. The phase of the field depends on the phase of the current at the instant the wave left the antenna. The strength of the magnetic component H of the wave is related to the electrostatic voltage gradient ϵ by the equation

$$\epsilon = 300H \tag{222}$$

where H is in lines per square centimeter and ϵ is in volts per centimeter.

The total field radiated from an antenna is found by adding up the separate fields produced by the elementary lengths of the radiator, taking into account phase relations and planes of polarization in making this addition. When the antenna configuration and the current distribution are known, the radiated field is determined by integrating the contributions that are made by each elementary length, as discussed in detail in Sec. 131.

The wave front of the radiated wave lies in a plane perpendicular to a line drawn toward the antenna, and the waves are polarized in the same direction as the antenna. Thus a plane can be passed through the antenna and an electrostatic flux line of the radiated wave, while the magnetic flux is perpendicular to such a plane.

Current Distribution in an Antenna.—An antenna represents a circuit having distributed constants, and so has a current distribution of the type discussed in Sec. 15. Strictly speaking, the inductance and capacity per unit length are not the same for all parts of the antenna, so that an exact solution for the current distribution is extremely complicated. Experiments have shown, however, that in the usual case where the antenna is operated so as to give resonance, the current is very nearly sinusoidally distributed, with the phase differing by 180° in adjacent half-wave-length sections (see Figs. 33 and 34).[1] The current is zero at the open ends of such an antenna and approaches zero at all points that are an exact multiple of a half wave length distant from the open end, while the current is maximum at points that are odd quarter wave lengths distant from the open ends. The length of an antenna expressed in wave lengths is very nearly equal to the length between extreme ends, measured in terms of the wave length of a wave traveling with the velocity of light, *i.e.*, a wave length of the radio wave. Examples of current distribution in a number of typical antennas are shown in Fig. 371. The current in

[1] This is shown by experimental results published by R. H. Wilmotte, Distribution of Current in a Transmitting Antenna, *Jour. I.E.E.* (*London*), vol. 66, p. 617, June, 1928. The only important practical exception is in the tower antennas used in broadcast work, where the variation in cross section with height modifies the current distribution (see Sec. 136).

each case follows a sinusoidal law, and is zero at the open ends. When the lower end of the antenna is grounded, as in a to f, or when the length of an ungrounded antenna is not an exact multiple of a half wave length as at j, the current distribution is made up of sections of sine waves as shown.

The common method of representing antenna current distribution, shown in Fig. 371, is not strictly correct, for, as explained in Sec. 15, losses prevent the current from going to zero except at an open end. The resulting error is so small, however, that it can always be neglected.

Impedance of Antennas:—An antenna, as a result of its distributed inductance and capacity, is equivalent to a transmission line and therefore, like all transmission lines, acts in many respects as a resonant circuit. In the case of an ungrounded antenna, resonance is obtained whenever the total length approximates a multiple of a half wave length, while with the grounded antenna resonance occurs whenever the length is an odd multiple of a quarter wave length. Consequently a voltage applied in

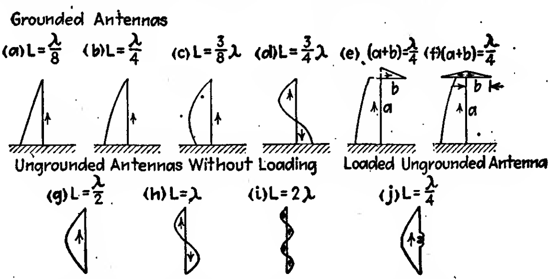


FIG. 371.—Current distribution in typical antennas. In each case the current has a sinusoidal distribution and is zero at the open ends.

series with an antenna will encounter a capacitive reactance at frequencies just below resonance, an inductive reactance at frequencies above resonance, and a resistance at resonance. If an antenna is not exactly the correct length for resonance, it can accordingly be brought into resonance by adding series inductance to neutralize the capacitive reactance of an antenna that is too short or by adding a series capacity to neutralize the inductive reactance of an antenna that is too long.

The magnitude of the impedance offered by an antenna to a series voltage depends upon the frequency, the length, antenna construction, ground losses, etc. Because of these many factors involved, the exact impedance cannot ordinarily be predicted, although fairly accurate qualitative estimates of behavior can be made.

Effect of the Ground; Image Antennas.—When an antenna is near the ground, energy radiated toward the earth is reflected as shown in Fig. 372, so that the total field in any direction represents the sum of a direct wave plus a reflected wave. In the case of a perfect earth (infinite conductivity) the reflection is complete, and can be taken into account by

replacing the ground by an image antenna as illustrated in Figs. 372 and 373. The fields produced by the joint action of the actual antenna and its image are the same as exist in the space above earth with the actual antenna in the presence of the perfect ground.

Examples of image antennas for a number of cases are given in Fig. 373. The general principles for setting up the image antenna for a perfect earth are as follows:

1. The image antenna has a physical configuration that is the mirror image of the actual antenna.

2. The currents in corresponding parts of the actual and image antennas (*i.e.*, parts lying on the same vertical line and at the same distance from the earth surface) are of the same magnitude and flow in the same direction when the corresponding parts are vertical and in the opposite direction when they are horizontal. The modifications necessary to take into account the effect of earth losses are considered in Sec. 132.

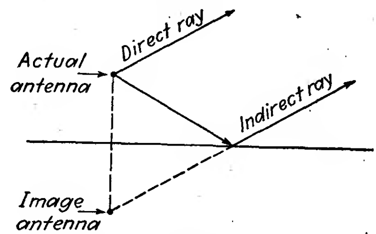


FIG. 372.—Diagram illustrating how the wave reflected from the earth can be considered to be produced by an image antenna.

Energy Relations and Antenna Resistance.—The rate at which energy passes through 1 sq cm of surface located in the wave front is the average amount of energy contained in 1 cc of wave multiplied by the velocity of light. Thus, if the wave at the point in question has a strength ϵ volts

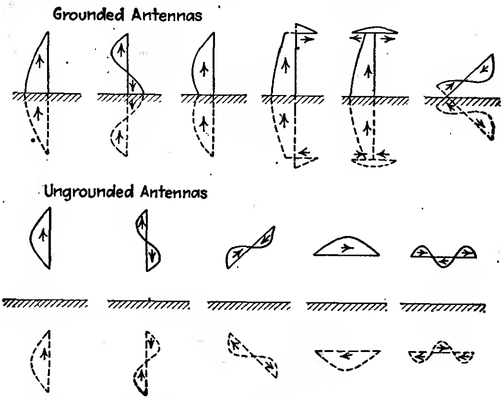


FIG. 373.—Images for common types of antennas.

effective value per centimeter, the average energy per cubic centimeter is $\epsilon^2 0.08842 \times 10^{-12}$ joule. Multiplying this by the velocity of light in centimeters per second shows that the rate at which energy flows through each square centimeter of wave front is $0.00265\epsilon^2$ joule per second. Hence the energy radiated from the entire antenna system can be determined by imagining that the antenna is at the center of a very large sphere and then adding up the energy that flows through each square

centimeter of the spherical surface that is above the ground. Practical means of evaluating this total energy are discussed in Sec. 134.

The total amount of energy radiated from a transmitting system can be conveniently measured in terms of a "radiation" resistance, which is the resistance that, when inserted in series with the antenna, will consume the same amount of power as is actually radiated. While the radiation resistance is a purely fictitious quantity, the antenna acts as though such a resistance were present because the loss of energy by radiation is in its effects equivalent to a like amount of energy dissipated in a resistance. The value of the radiation resistance is determined by the antenna construction, and particularly the size measured in wave lengths, the relation to ground and other conducting objects, such as towers, buildings, and trees, etc., and by the point on the antenna to which the resistance is referred (*i.e.*, the point at which the radiation resistance is considered as being lumped). Unless specifically stated otherwise, it is customary to refer the radiation resistance to a current loop. Methods of evaluating the radiation resistance are taken up in Sec. 134.

In addition to the radiated energy, energy is also lost in the antenna system as a result of wire and ground resistance, corona, eddy currents induced in neighboring masts, guy wires and other conductors, and dielectric losses arising from imperfect dielectrics, such as trees and insulators, located in the field of the antenna. These losses can be represented in the same way as the radiated energy, *i.e.*, by a resistance which when inserted in series with the antenna will consume the same amount of power as is actually dissipated in these various ways. The total antenna resistance is the sum, $R_r + R_l$, of the radiation resistance R_r and the loss resistance R_l, and determines the amount of energy that must be supplied to the antenna to produce a given current.

The efficiency of the antenna as a radiator is the ratio $R_r/(R_r + R_l)$ of radiation to total resistance. This represents the fraction of the total energy supplied to the antenna which is converted into radio waves.

Effective Height.—The term *effective height* as applied to a transmitting antenna represents the length of elementary antenna of Fig. 370 which, when carrying a uniform current equal to the current flowing at the reference point in the actual antenna, will produce the same field intensity as is actually radiated. The effective height is very seldom used in connection with transmitting, but is a convenient conception when the antenna acts to receive radio waves.

Induction Fields.[1]—The electrostatic and magnetic fields having strengths given by Eqs. (221) and (222) do not include the ordinary

[1] For a more complete consideration of the fields existing near an antenna, see F. R. Stansel, A Study of the Electromagnetic Field in the Vicinity of a Radiator, *Proc. I.R.E.*, vol. 24, p. 802, May, 1936; P. S. Carter, Circuit Relations in Radiating Systems, *Proc. I.R.E.*, vol. 20, p. 1004, June, 1932.

magnetic and electrostatic induction fields that are present in the immediate vicinity of the antenna even at low frequencies where the radiation is negligible. The total magnetic field that is produced in the vicinity of an elementary antenna, such as that of Fig. 370, is given by the following equation:

Magnetic field in lines per square centimeter =

$$\frac{2\pi(\delta l)I}{10d\lambda}\cos\omega\left(t-\frac{d}{c}\right)\cos\theta-\left(\frac{(\delta l)I}{10d^2}\right)\sin\omega\left(t-\frac{d}{c}\right)\cos\theta \quad (223)$$

The notation is the same as in Eq. (221) except that all the units of length are in centimeters. The first term of this equation represents the radiation field given in Eq. (222), while the second term represents the induction field that gives rise to the self-inductance of the antenna system. The induction magnetic field is in phase with the current flowing in the radiator (after making allowance for the time required in propagation) and is inversely proportional to the square of the distance, while the radiation field is in time quadrature, *i.e.*, 90° out of phase, with the current and is inversely proportional to the first power of the distance. Because of the way in which it varies with distance, the induction field is of importance only in the immediate vicinity of the antenna, where it is much stronger than the radiation field. At a distance of $\lambda/2\pi$ from the antenna the two fields are equal, while at greater distances the radiation field predominates.

Along with the induction magnetic field there is also an induction electrostatic field, which is in time phase with the electrostatic field of the radiated wave and, like the magnetic induction field, dies out much more rapidly with distance than does the radiated wave.

130. Fundamental Properties of Receiving Antennas and Reciprocal Relations Existing between Transmitting and Receiving Properties.—A receiving antenna is able to abstract energy from a passing radio wave as a result of the voltages that the magnetic flux of the wave induces in the antenna. These induced voltages are distributed along the entire length of the antenna and have a value which per meter of antenna length is $\epsilon \cos \psi \cos \theta$, where ϵ is the field strength of the wave in volts per meter, ψ is the angle between the plane of polarization and the wire in which the voltage is induced, and θ is the angle between the wave front and the direction of the antenna wire. It will be observed that the quantity $\epsilon \cos \psi \cos \theta$ is the component of the field strength which has a wave front parallel to the antenna and is polarized in the same plane as the antenna.

Energy and Current Delivered to Load Impedance.—In determining the current that the distributed induced voltages will develop in a load impedance inserted in series with the antenna, it is convenient to replace the actual antenna with its load by the equivalent circuit of Fig. 374. Here the antenna is considered as a generator consisting of an equivalent lumped

voltage E having an internal impedance Z_a with which the load impedance Z_L is in series. The ratio of lumped voltage to the strength of the radio wave is termed the effective height h of the receiving antenna (*i.e.*, $E = \epsilon h$), while the equivalent impedance Z_a is the impedance that the load sees when looking toward the antenna. The current that flows in this equivalent circuit is exactly the same as the current that flows in the corresponding part of the real antenna.

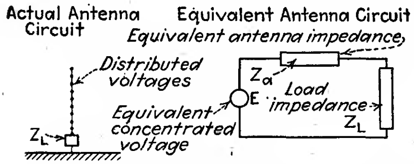


FIG. 374.—Actual receiving antenna with load impedance Z_L and distributed induced voltages, together with equivalent antenna circuit in which the distributed voltages are replaced by a lumped voltage E and the distributed antenna impedance by an equivalent concentrated impedance Z_a.

The energy absorbed by the load is maximum when the resistance of the load equals the resistance component of the antenna impedance and the reactance of the load is equal in magnitude but opposite in sign to the reactive component of the equivalent antenna impedance Z_a. When the antenna is tuned to resonance with the frequency being received, the antenna impedance Z_a is a pure resistance equal to the total antenna resistance, $R_r + R_l$, referred to the point where the load resistance is to be inserted, and the proper value of load resistance is $R_L = (R_r + R_l)$.

The total energy that the receiving antenna abstracts from the passing radio wave represents the energy dissipated in the equivalent antenna circuit, and so is given by the relation

$$\text{Total power in watts abstracted from radio wave} = \frac{(\epsilon h)^2}{R_L + R_r + R_l} \tag{224}$$

where

ϵ = field strength (r.m.s. value) of the radio wave in volts per meter
h = effective height of the antenna in meters
R_r = antenna radiation resistance
R_l = antenna loss resistance
R_L = load resistance.

The fraction $R_L/(R_r + R_l + R_L)$ of this total energy represents the portion of the abstracted energy which is usefully employed. Of the ramainder, part is accounted for by the antenna losses, such as wire and ground resistance, etc., while the rest is reradiated. This reradiation of energy results from the fact that, when current flows in an antenna, radiation takes place irrespective of whether the voltage producing the current is derived from a passing radio wave or from a vacuum tube.

The maximum amount of energy that it is theoretically possible for a given antenna to abstract from a passing radio wave occurs when

the antenna loss resistance is made negligibly small and when the load resistance R_L is equal to the radiation resistance. Under these conditions the rate at which energy is abstracted from the wave is $(\epsilon h)^2/2R_r$ watts, and, since half of this is reradiated, the maximum possible amount of power that can be delivered to a load resistance R_L is $(\epsilon h)^2/4R_L$ watts. Calculations show that a section of wave front extending for only about one-quarter of a wave length on each side of the receiving antenna will be capable of supplying this received energy.

The electromagnetic and electrostatic fields in the vicinity of a receiving antenna are the sum of the fields produced by the radio wave and by the current in the receiving antenna. The result is that the receiving antenna causes a distortion in the field pattern in its immediate vicinity.[1] Analysis shows that the effect of the receiving antenna on the passing wave is, first, abstraction of energy, which weakens the main wave, and, second, reflection or reradiation of energy, which redistributes the energy of the passing wave in a manner depending upon the antenna tuning (see Sec. 135 for a further discussion of this point).

Relations between Receiving and Radiating Properties and the Rayleigh-Carson Reciprocity Theorem.—The properties of an antenna when used to abstract energy from a passing radio wave are similar in nearly all respects to the corresponding properties of the same antenna when acting as a radiator. Thus the directional characteristics, the current distribution, the effective height, and the impedance of the antenna are the same in reception as in transmission. The only difference is in the radiation resistance, which in the case of receiving antennas depends upon the inserted load impedance and tends to be higher than when radiating.[2] These reciprocal relations between the transmission and reception properties are extremely useful because they make it possible to deduce the merits of a receiving antenna from transmission tests, and vice versa. Thus the directional characteristics of a given antenna can be determined either by operating this antenna as a receiver and moving a portable transmitter about or by using the antenna as a transmitter and making observations of field strength on a portable receiving set. Similarly the effective height can be calculated by determining the voltage that must be inserted in the antenna to produce the same current as is produced by a wave of known strength, or it can be obtained by determining the field strength that the antenna produces at a known distance when the antenna is transmitting with a known current.

[1] See Henry C. Forbes, Re-radiation from Tuned Antenna Systems, *Proc. I.R.E.*, vol. 13, p. 363, June, 1925; F. A. Kolster and F. W. Dunmore, The Radio Direction Finder and Its Application to Navigation, *Bur. Standards Sci. Paper* 428.

[2] See Raymond M. Wilmotte, Generalized Theory of Antennae, *Exp. Wireless and Wireless Eng.*, vol. 5, p. 119, March, 1928.

These reciprocal relations between the transmitting and receiving properties of an antenna are incorporated in reciprocal theorems, the most important of which was discovered by Rayleigh and extended to include radio communication by John R. Carson.[1] It is to the effect that, *if an electromotive force E inserted in antenna 1 causes a current I to flow at a certain point in a second antenna 2, then the voltage E applied at this point in the second antenna will produce the same current I (both in magnitude and phase) at the point in antenna 1 where the voltage E was originally applied.* The Rayleigh-Carson theorem fails to be true only when the propagation of the radio waves is appreciably affected by an ionized medium in the presence of a magnetic field, and so holds for all conditions except short-wave transmission over long distances. Even then, it is to

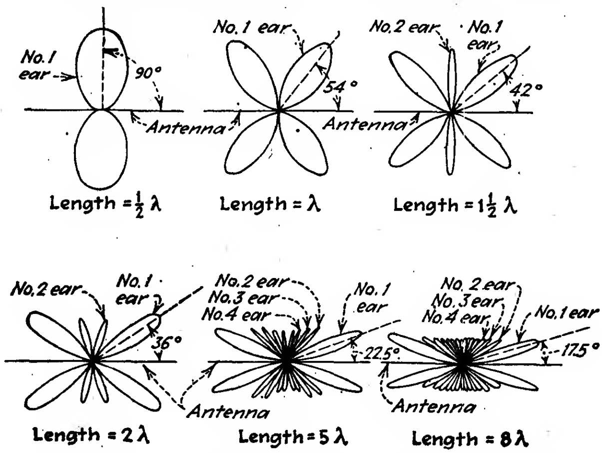


FIG. 375.—Polar diagrams showing strength of field radiated in various directions from an antenna consisting of a wire remote from the ground.

be expected that *on the average* the theorem will still apply, even when it cannot be depended upon to be exactly correct in every instance.

131. Directional Characteristics of Simple Antennas.—The directional characteristics of transmitting antennas are important because only those waves radiated in certain directions from the transmitter will reach a particular receiving point, while all energy radiated in other directions is wasted as far as transmission to this receiver is concerned. The directional characteristics of receiving antennas are likewise important because an antenna that abstracts much larger quantities of energy from waves coming from the direction of the transmitter than from waves of equal strength arriving from other directions will not pick up interfering signals and static arriving from other directions.

[1] See John R. Carson, Reciprocal Theorems in Radio Communication, *Proc. I.R.E.*, vol. 17, p. 952, June, 1929; John R. Carson, The Reciprocal Energy Theorem, *Bell System Tech. Jour.*, vol. 9, p. 325, April, 1930.

The fundamental factors determining the directive characteristics of antennas can be understood by considering a series of simple cases which incorporate the various principles involved.

Wire Remote from Ground.—The directional characteristics of an antenna consisting of a single wire having a length that is an exact multiple of a half wave length (*i.e.*, an antenna operated at a resonant frequency) and far removed from other objects (particularly the ground) is given by Eq. (238). Polar diagrams giving the directional characteristics of several such antennas in a plane containing the wire are shown in Fig. 375. If the directional characteristic is viewed in three-dimensional space, it consists of a figure of revolution having a cross section as shown in Fig. 375. An examination of Fig. 375 shows that the directional characteristic contains a number of lobes, the largest of which are always the lobes making the smallest angle with the axis of the wire. Increasing the antenna length as measured in wave lengths has the effect of reducing the angle between the axis of the wire and the direction of the large lobes, and also of increasing the number of lobes present, as is clearly evident in the figure.

The important features of the directional characteristics of a wire antenna in space are incorporated in the curves of Fig. 376, which show the angles with respect to the wire axis at which the radiation is maximum and zero, and also show the relative amplitude of the successive lobes or ears. Thus reference to Fig. 376 shows that a wire five wave lengths long has five lobes in each quadrant, with the maximum of these lobes coming at angles of 22.5°, 46°, 60°, 73°, and 84° with respect to the wire axis, while the relative amplitudes of these lobes are 2.25, 1.40, 1.20, 1.05 and 0.95, respectively.

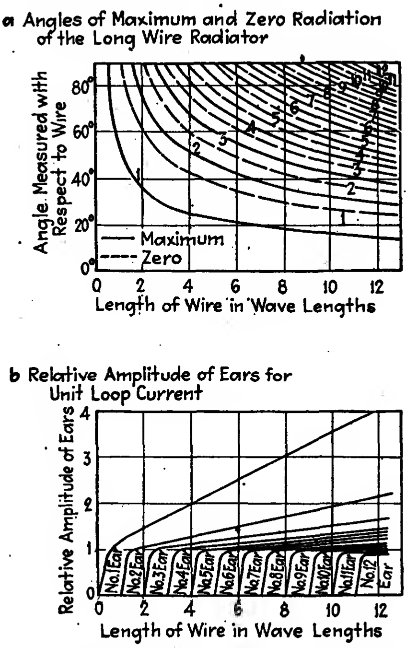


FIG. 376.—Angles at which the radiation from an isolated wire antenna is zero and maximum, together with relative field strength of ears. The ears are numbered so that ear 1 is the lobe nearest the direction in which the antenna points.

The reason that the directional characteristic of a long wire differs from that of an elementary antenna is that the current in different parts of the long antenna may not be in the same phase, and the distance from a remote point to various parts of the long antenna will not be the

same. The result is that the fields radiated from different portions of the long antenna add together vectorially and give a sum that depends upon the direction and upon the current distribution along the wire. This can be made clear by referring to Fig. 377 and considering the radiation produced at distant points P, P_1, and P_2 by elementary sections located at x and $-x$ on the antenna. The observer at P is equidistant from both points under consideration, and, since the radiation at x is in opposite phase from the radiation at $-x$, the two cancel at P to give zero field. In contrast with this, the observer at point P_1 is closer to x than to $-x$, so that, although the waves start from these two points in phase opposition, they do not completely cancel at P_1. Finally at point P_2 there is no field because an elementary length of the antenna does not radiate energy along its axis. The result is a directive pattern in which the maximum of radiation occurs at an angle with respect to the wire axis, as shown.[1]

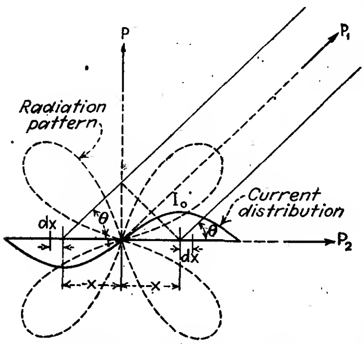


FIG. 377.—Diagram illustrating factors controlling the directional characteristic of an antenna. Note that the radiation does not cancel in the direction P_1 because of the different distances to the two ends of the antenna.

Grounded Vertical Antenna.—Typical examples of grounded antennas are illustrated in the top lines of Figs. 371 and 373. When the height of a grounded antenna is of the order of one-eighth wave length or less, the radiation from the vertical portion is almost exactly proportional to the cosine of the angle of elevation, and the directional characteristic is as shown in Fig. 378 (assuming a perfect earth). This is because the total length of the antenna plus its image is then so small that to a distant observer the difference in distance to the different parts of the antenna is a small fraction of a wave length. The directional characteristic is hence substantially that of an elementary antenna.

[1] The directive pattern obtained from a long wire can be readily calculated by considering the antenna to be made up of a number of elementary lengths and then adding up the resulting fields produced at a distant point as indicated in Sec. 129. Thus consider the case where the antenna is an even number of half wave lengths long, as is the case in Fig. 377, and assume that the current is sinusoidally distributed. If the reference point is the midpoint of the antenna ($x = 0$ midpoint) and I_0 is the maximum current, then the equation of the current distribution can be written as

$$i = I_0\left(\sin \frac{2\pi x}{\lambda}\right)[\cos (\omega t + 90°)] \qquad (225)$$

Radiation at a distance x from the midpoint reaches distant point P_1 sooner than radiation from the reference point because the distance is less by $x \cos \theta$, which

When the antenna has a flat top as in the case of the two antennas shown at the end of the top line of Fig. 371, the radiation from the flat top will be substantially zero unless the antenna height approaches or exceeds a quarter wave length. This is because with low heights the radiation from the flat top is almost completely neutralized by the radiation from the image of the flat top.

When the grounded antenna consists of a vertical wire that is a half wave length or more in length, the directional characteristics in the vertical plane are markedly affected, as is apparent from Fig. 378. These patterns come about as a result of the fact that such an antenna, together with its

---

corresponds to $(2\pi x/\lambda)\cos\theta$ radians advance in phase. By combining Eqs. (221) and (225), the field $\Delta\epsilon$ produced at P_1 by an elementary length of the antenna is:

$$\Delta\epsilon = \frac{60\pi}{d\lambda} I_0 \sin\frac{2\pi x}{\lambda}\cos\left(\omega t - \frac{d_0}{c} + \frac{2\pi x}{\lambda}\cos\theta\right)\sin\theta\,dx \tag{226}$$

where

θ = angle of elevation measured with respect to the wire axis

c = velocity of light

d_0 = distance from the wire center to P_1.

The total radiation from the entire length L of the antenna is

$$\begin{aligned}
\epsilon &= \frac{60\pi}{d\lambda} I_0 \sin\theta \int_{x=-\frac{L}{2}}^{x=\frac{L}{2}} \sin\frac{2\pi x}{\lambda}\cos\left(\omega t - \frac{d_0}{c} + \frac{2\pi x}{\lambda}\cos\theta\right)dx \\
&= \frac{60\pi}{d\lambda} I_0 \sin\theta \int_{x=-\frac{L}{2}}^{x=\frac{L}{2}} \sin\frac{2\pi x}{\lambda}\left[\cos\left(\frac{2\pi x}{\lambda}\cos\theta\right)\cos\left(\omega t - \frac{d_0}{c}\right)\right. \\
&\qquad \left. - \sin\left(\frac{2\pi x}{\lambda}\cos\theta\right)\sin\left(\omega t - \frac{d_0}{c}\right)\right]dx \qquad (227) \\
&= \frac{60\pi}{d\lambda} I_0 \sin\theta \left\{\cos\left(\omega t - \frac{d_0}{c}\right)\left\{\frac{-\cos\left[\frac{2\pi x}{\lambda}(1-\cos\theta)\right]}{\frac{4\pi}{\lambda}(1-\cos\theta)} - \frac{\cos\left[\frac{2\pi x}{\lambda}(1+\cos\theta)\right]}{\frac{4\pi}{\lambda}(1+\cos\theta)}\right\}\right. \\
&\qquad \left. - \sin\left(\omega t - \frac{d_0}{c}\right)\left\{\frac{\sin\left[\frac{2\pi x}{\lambda}(1-\cos\theta)\right]}{\frac{4\pi}{\lambda}(1-\cos\theta)} - \frac{\sin\left[\frac{2\pi x}{\lambda}(1+\cos\theta)\right]}{\frac{4\pi}{\lambda}(1+\cos\theta)}\right\}\right\}_{-\frac{L}{2}}^{\frac{L}{2}}
\end{aligned}$$

When the length is a whole number of wave lengths, substitution of the limits, followed by a few transformations, gives:

$$\epsilon = \frac{60}{d} I_0 \frac{\sin\left(\frac{\pi L}{\lambda}\cos\theta\right)}{\sin\theta}(-1)^n \sin\left(\omega t - \frac{d_0}{c}\right) \tag{228}$$

where n is the number of wave lengths.

The same procedure is followed when the antenna length is an odd number of half wave lengths, except that the current distribution is now $i = I_0\cos(2\pi x/\lambda)\cos(\omega t + 90°)$. The directivity in this case is given by Eq. (238a).

image, is so long that the distance from different parts of the antenna to a distant observer may differ by an appreciable fraction of a wave length for the higher vertical angles.

Spaced Antennas.—One of the most important means of obtaining directivity is by combining the radiation from two or more spaced antennas. A simple example of such directivity is illustrated in Fig. 379, which shows two vertical antennas spaced a quarter of a wave length apart, carrying equal currents and excited so that they are 90° out of phase. To an observer at the point P_3 the radiation from the two anten-

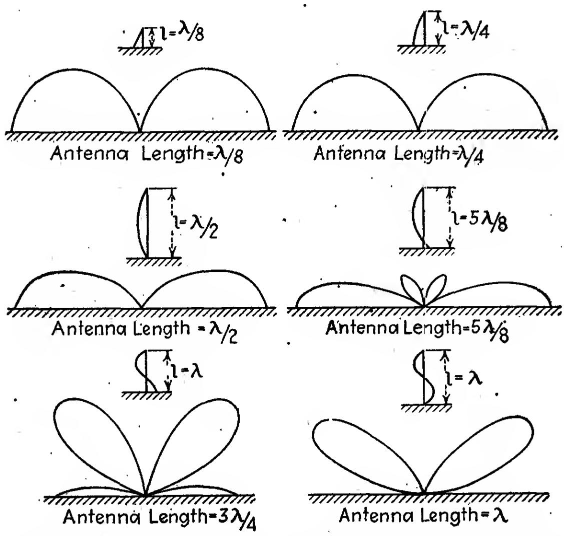


FIG. 378.—Directional characteristics in a vertical plane of field produced by grounded vertical antennas of varying lengths. These polar diagrams can be thought of as cross sections of a figure of revolution about the axis of the antenna, and assume a perfectly conducting earth.

nas adds because, although the radiation leaves antenna A a quarter of a cycle ahead of the radiation from antenna B, it takes this first radiation a quarter of a cycle to travel the quarter wave length to antenna B, and the two radiations add in the direction toward P_3. However, to an observer at P_1 the radiations from the two antennas cancel because the radiation from antenna B starts a quarter cycle late and loses an additional quarter cycle in traveling to A, thereby arriving exactly in the correct phase to cancel the radiation that is starting out from A towards P_1. A distant observer at P_2 is equidistant from the two antennas, and, since the radiations start out with a phase difference of 90°, they will add up in quadrature at P_2 to give a result that is $\sqrt{2}$ times the radiated field from a single antenna. The resulting directional pattern is a cardiod as shown in Fig.

379.[1] By changing either the spacing, the relative phase, or both, other directional patterns are obtained as illustrated in Fig. 380.

More complicated systems of spaced antennas are considered in the section below on Antenna Arrays.

Non-resonant Long-wire Antennas. When the end remote from the source of power is properly terminated, a transmission line and hence an antenna have current distributions such as illustrated in Fig. 33*c*. The current dies away exponentially with distance from the sending end, and there is a phase shift that increases uniformly with distance and amounts to 360° per wave length. The equation of the current distribution is hence

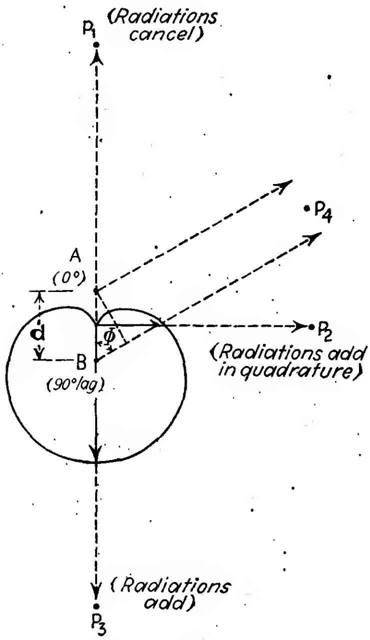


FIG. 379.—Diagram illustrating how directivity is obtained by the use of two spaced antennas. In the particular case shown the spacing is $\lambda/4$, the currents are equal, and the relative phase is 90°.

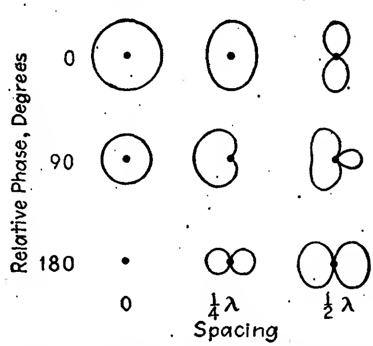


FIG. 380.—Directive patterns obtained from a pair of spaced antennas with different spacing and phasing, but carrying currents of equal magnitude.

$$\left.\begin{array}{l}\text{Current at distance } x\\ \text{from power source}\end{array}\right\} = I_0\epsilon^{-\alpha x}\cos\left(\omega t + 90° - \frac{2\pi x}{\lambda}\right) \qquad (230)$$

[1] The equation of the field pattern obtained from two spaced antennas can be readily derived as follows: To a distant observer at P_4 antenna A is closer by $d\cos\Phi$, where d is the antenna spacing and Φ is the angle, as shown in Fig. 379. If the phase of the currents in B lags by α radians behind A, then the phase difference between the radiations arriving from A and B at P is $(2\pi d/\lambda)\cos\Phi + \alpha$ radians. The relative field at P as a function of the angle Φ is then

$$\text{Relative field} = 2\cos\left[\left(\frac{\pi d}{\lambda}\right)\cos\Phi + \frac{\alpha}{2}\right] \qquad (229)$$

When $d = \lambda/4$ and $\alpha = \pi/2$, this becomes a cardiod as shown in Fig. 380.

where

$I_0 \cos(\omega t + 90°)$ = current entering the line at the sending end

α = attenuation factor determined by losses (including radiation)

λ = distance corresponding to a wave length.

A long wire with such a current distribution has a directional characteristic determined by the same factors that have been considered above, but, because the law of amplitude and phase variation of the current are different, the directional characteristics will as a result be modified. The results for several typical cases are shown in Fig. 381.[1] Here it is seen that the maximum radiation takes place at an angle with respect to the axis of the wire which decreases as the wire becomes longer and is very little affected by even considerable attenuation. Attenuation modifies the shape of the pattern in various details, however. It will be noted that in all cases the radiation pattern is substantially unidirectional.

Gain of Directive Antennas.—The merit of a directive antenna is most conveniently measured in terms of the antenna *gain*, which is defined as

[1] The formula from which these patterns are calculated is derived as follows: Referring to Fig. 381, the current at x lags the current at the generator by $2\pi x/\lambda$ radians, but is closer to P by $x \cos \theta$, which corresponds to $(2\pi x/\lambda) \cos \theta$ radians, so that, relative to the radiation from the generator end, the radiation arriving at P from dx lags by $(2\pi x/\lambda)(1 - \cos \theta)$. From Eq. (221), the radiation $\Delta\epsilon$ from dx is

$$\Delta\epsilon = \frac{60\pi}{d\lambda}(\sin \theta) I_0 \epsilon^{-\alpha x} \sin\left[\omega t - \frac{2\pi x}{\lambda}(1 - \cos \theta)\right] dx \qquad (231)$$

Accordingly, the total radiation from the entire length L is

$$\begin{aligned}\epsilon &= \frac{60\pi}{d\lambda} I_0 \sin \theta \int_0^L \epsilon^{-\alpha x} \sin\left[\omega t - \frac{2\pi x}{\lambda}(1 - \cos \theta)\right] dx \\ &= \frac{60\pi}{d\lambda} I_0 \sin \theta \int_0^L \epsilon^{-\alpha x} \left\{\sin \omega t \cos\left[\frac{2\pi x}{\lambda}(1 - \cos \theta)\right]\right. \\ &\qquad \left. - \cos \omega t \sin\left[\frac{2\pi x}{\lambda}(1 - \cos \theta)\right]\right\} dx \qquad (232)\end{aligned}$$

Upon integrating and simplifying, this becomes

$$\epsilon = \frac{k}{\sqrt{\alpha^2 + p^2}} \epsilon^{-\alpha x} \cos\left[\omega t - (px - \Phi)\right]\Big|_0^L$$

where

$$k = \frac{60\pi}{d\lambda} I_0 \sin \theta$$

$$p = \frac{2\pi}{\lambda}(1 - \cos \theta)$$

$$\Phi = \arctan \frac{\alpha}{p}.$$

Substituting the limits and making further simplifications then yields

$$\epsilon = \frac{30 I_0 \sin \theta}{d\sqrt{1 + (\alpha/p)^2}(1 - \cos \theta)}(1 + \epsilon^{-2\alpha L} - 2\epsilon^{-\alpha L} \cos pL)^{1/2} \qquad (233)$$

the power that must be supplied to a standard comparison antenna to lay down a given field strength in the desired direction, divided by the power that must be supplied to the directive antenna to accomplish the same result. Thus a gain of 100 (=20 db) means that the directive antenna requires only $\frac{1}{100}$ as much power to produce a given field strength in the desired direction as does the comparison antenna; or, conversely, when both antennas are supplied with the same power, the directive system will make the received energy 100 times as great (corresponding to a tenfold increase in field strength).

The comparison antenna is usually taken either as a wire one-half wave length long and at arbitrary orientation and height above earth or as one of the individual antennas of which the array is composed. The latter

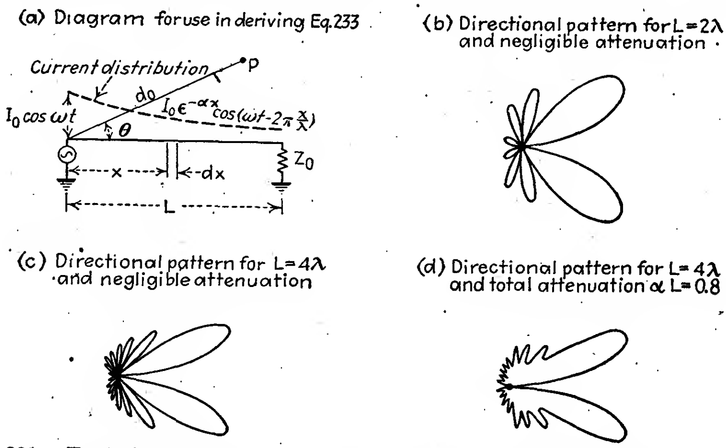


FIG. 381.—Typical directional characteristics for a non-resonant wire in space. These patterns may be considered as cross sections of a figure of revolution in which the wire is the axis.

method of expressing gain has the merit of giving a result that is nearly independent of earth effects and of the directive characteristics of the comparison antenna, except when the comparison antenna is sharply directional in the same planes as is the array.

Methods of calculating gain are discussed in Sec. 133.

132. Antenna Arrays.—Directional arrangements involving a combination of two or more spaced antennas are termed antenna arrays. The array represents one of the most widely used methods of obtaining directivity, and is accordingly of such importance as to warrant special consideration. The principal characteristics of the more important types of arrays are discussed below.[1]

[1] The discussion on broadside, end-fire, and colinear arrays and also arrays of arrays is based largely upon the following references: Ronald M. Foster, Directive Diagrams of Antenna Arrays, *Bell System Tech. Jour.*, vol. 5, p. 292, April, 1926;

Broadside Array.—A broadside array consists of a number of antennas uniformly spaced along a line, carrying equal currents and all excited in the same phase. The broadside array is characterized by a concentration of radiation in a plane at right angles to the line of the array, with a tendency for radiation in other directions to be canceled.

The principal properties of a broadside array are illustrated in Figs. 382 and 383. The sharpness of the beam in the horizontal plane increases with the length of the array and is substantially independent of the spacing of the radiators until a critical spacing is exceeded, which in the case

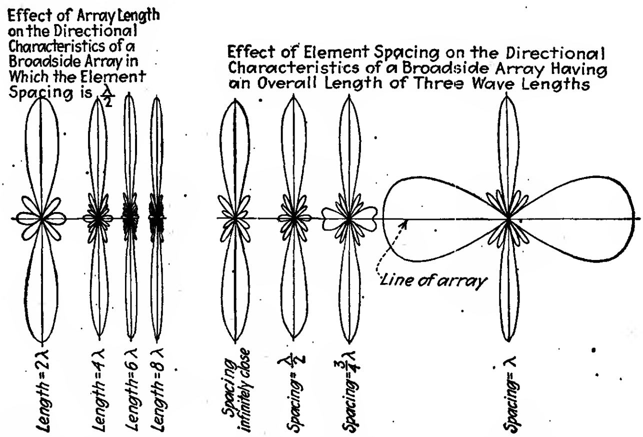


FIG. 382.—Effect of array length and element spacing on the directional characteristics of field radiated from a broadside array in a horizontal plane.

of Fig. 382 is $3\lambda/4$. The directivity in a vertical plane at right angles to the line of the array is the same as the directivity of the individual radiators, while in other vertical planes there is a tendency for the radiation to be either partially or completely canceled.

The gain of a broadside array consisting of short vertical antennas is shown in Fig. 383; it is almost exactly proportional to the length of the array and substantially independent of the spacing up to the critical spacing. It will be observed that the gain realized is large if the array is long.

The individual antennas making up the array can be of any desired type. Thus it is possible to use horizontal radiators instead of vertical radiators, or each elementary radiator of the array can in itself be an array, thus giving an array of arrays. In any case, the broadside arrangement increases the sharpness of the beam in a plane containing the line of the array.

G. C. Southworth, Certain Factors Affecting the Gain of Directive Antennas, *Proc. I.R.E.*, vol. 18, p. 1502, September, 1930; E. J. Sterba, Theoretical and Practical Aspects of Directional Transmitting Systems, *Proc. I.R.E.*, vol. 19, p. 1184, July, 1931.

The maximum permissible spacing of the individual radiators within the array depends upon the extent to which the individual radiators concentrate their radiation in the desired direction. If the elementary radiators have a nondirectional characteristic in the plane being considered, such as is the case in Fig. 382, then the maximum spacing is $3\lambda/4$; but if the individual radiators already possess some directivity, the spacing can be increased. Thus in a broadside array composed of horizontal radiators the antennas may be spaced as much as one wave length without spoiling the directional characteristic in the horizontal plane, because a horizontal wire possesses appreciable directivity in the horizontal plane. When the elementary radiators are themselves arrays with considerable horizontal directivity, spacings of two to three wave lengths are often permissible.

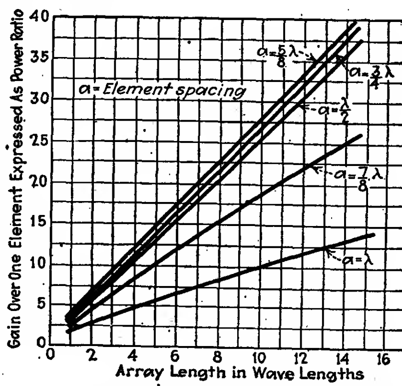


FIG. 383.—Gain of broadside array as a function of array length for various element spacings. These curves are for the case where radiation from the individual antenna is proportional to the cosine of the angle of elevation and give the gain over a single element of the array.

The number and character of the minor lobes in the directional characteristic of a broadside array depend upon the length of the array and the number of radiators contained. The longer the array the more numerous but also the smaller these minor lobes will be. If the spacing between radiators exceeds the critical value, one or more of the minor lobes will become so greatly enlarged as to spoil the directional characteristic, as illustrated in Fig. 382 for spacing $= \lambda$.

The directional characteristics of a broadside array can be calculated by summing up the fields produced by each element of each component of the array. Formulas for doing this are given in Sec. 133.

Colinear or Stacked Array.—This is a special form of broadside array where the line of the array is vertical. Such an arrangement gives directivity in a vertical plane, *i.e.*, in a plane containing the line of the array, and represents a common means of obtaining vertical directivity.

Typical directional characteristics under different conditions are shown in Fig. 384. The sharpness of the beam is proportional to the length of the array and is independent of element spacing up to a critical spacing, which in the case illustrated is one wave length. This critical spacing is determined by the same factors as in a broadside array, and would, for example, be $3\lambda/4$ if the radiators were horizontal wires.

End-fire Array.—The end-fire array consists of a series of antennas arranged in a line, carrying equal currents and excited so that there is a

progressive phase difference between adjacent antennas equal in cycles to the spacing between antennas in wave lengths. Thus, if the adjacent

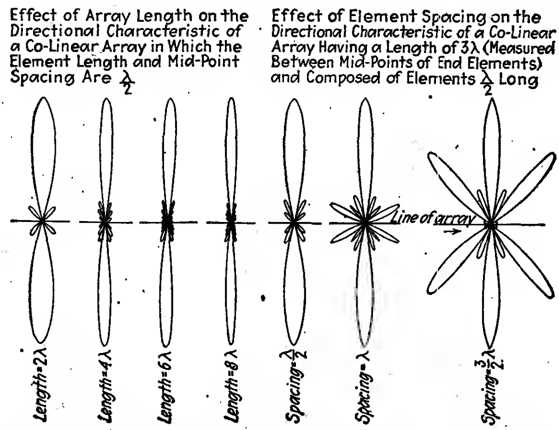


FIG. 384.—Effect of array length and element spacing on the directional characteristic of field radiated from a co-linear array in a plane containing the array.

antennas are a quarter of a wave length apart, a phase difference of 90° between adjacent antennas is called for. The result of such a phasing

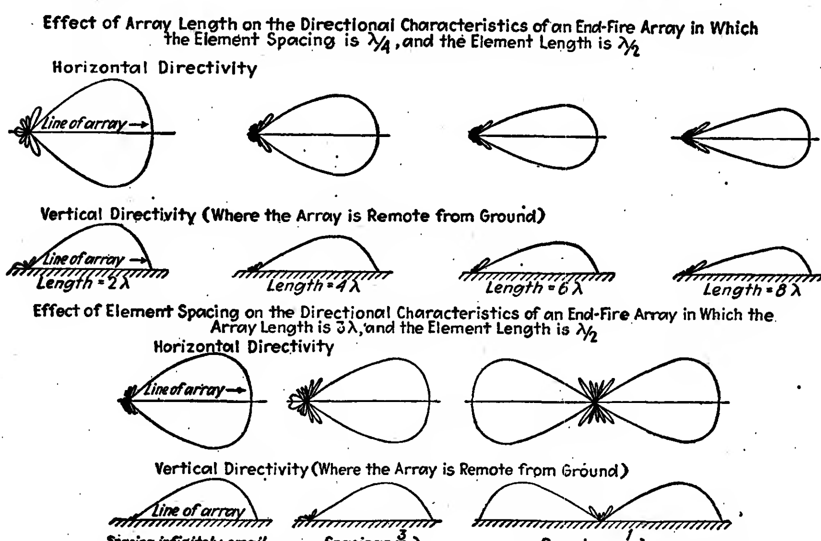


FIG. 385.—Effect of array length and element spacing on the directional characteristic of field radiated from an end-fire array.

system is to concentrate the radiation in the direction toward the end of the array having the most lagging phase, as illustrated in Fig. 385. The

unidirectional nature of the resulting characteristic makes the end-fire array of considerable practical importance.

The principal properties of an end-fire array are illustrated in Figs. 385 and 386. The degree of concentration in both horizontal and vertical planes is proportional to the length of the array, and is also independent of the spacing of the elementary radiators provided this spacing does not exceed a critical value. The gain of the array is likewise proportional to length and nearly independent of spacing up to the critical spacing. This critical spacing depends upon the directivity of the individual radiators and is $3\lambda/8$ with antennas whose length does not exceed $\lambda/2$, but it can be greater if the individual radiators have greater directivity.

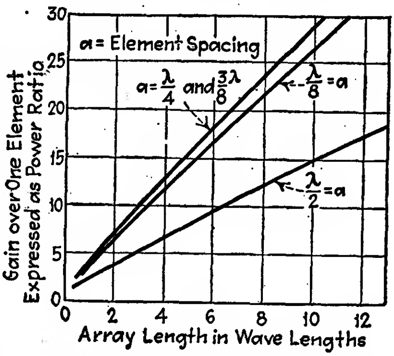


FIG. 386.—Gain of end-fire array as a function of array length for various element spacings. These curves are calculated on the assumption that the radiation from the individual antenna is proportional to the cosine of the angle of elevation.

The directive characteristics of end-fire arrays can be calculated in the same manner as the broadside array, by adding up the radiations produced in the desired direction by each element of each antenna. The necessary formulas for doing this are given in Sec. 133.

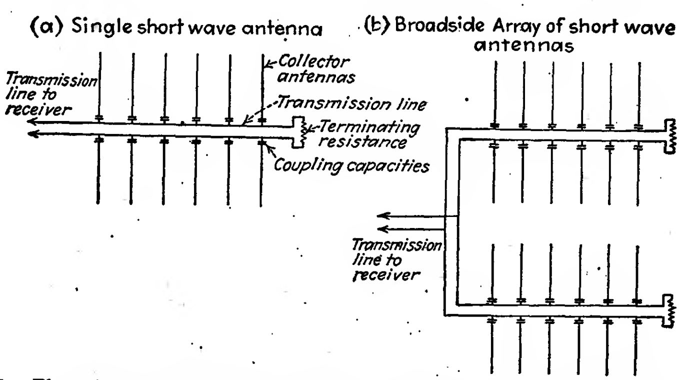


FIG. 387.—Plan view of single fishbone antenna and of an array consisting of two such antennas in broadside.

Fishbone Antenna.[1]—The fishbone antenna is illustrated in Fig. 387 and is a special form of end-fire array in which the necessary phase

[1] This type of antenna was developed by H. O. Peterson, and is described by H. H. Beverage and H. O. Peterson, Diversity Receiving System of R.C.A. Communications, Inc., for Radio Telegraphy, *Proc. I.R.E.*, vol. 19, p. 531, April, 1931.

relations are maintained by a non-resonant line. The outstanding characteristic of the fishbone antenna is the ability to perform satisfactorily over a considerable band of frequencies without any readjustments whatsoever.

The fishbone antenna consists of a series of collectors arranged in colinear pairs and loosely coupled to the transmission line by small capacities (usually the insulator capacity) as indicated in Fig. 387. The collectors are usually but not necessarily horizontal. The phasing is obtained by taking advantage of the fact that, when a transmission line is terminated in its characteristic impedance, there is a uniform phase shift along the line of 360° per wave length, which automatically phases the collector antennas properly with respect to each other.

In the usual design the collectors are each about 0.3 wave length long at the optimum frequency and are spaced a distance not to exceed $\lambda/12$ at this frequency. The length of the array is commonly three to five wave lengths, and is limited by the reactive loading that the coupled antennas place upon the transmission line. With these proportions the antenna is effective for frequencies ranging from about 1.2 to 0.5 of the optimum frequency, or over a frequency range exceeding 2 to 1.[1]

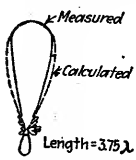


FIG. 388.—Directional characteristics in a horizontal plane of typical short-wave fishbone antenna.

A typical directional pattern of a fishbone antenna is given in Fig. 388, and is a unidirectional pattern characterized by very small minor lobes. The characteristics of the back-end radiation can be controlled to some extent by causing either the phase or magnitude of the terminating impedance to depart slightly from the characteristic impedance; and it is even possible in this way to make a null occur in almost any arbitrary backward direction. When the directivity desired in a horizontal plane is greater than that given by a single fishbone, two such arrays can be placed broadside as illustrated in Fig. 387*b*, thus giving a two-element broadside array, each component of which is a fishbone antenna.

Array of Arrays.—The antenna arrays used in practice are nearly always combinations that can be considered as an array, each element of which is itself an antenna array. The directional characteristic of such an array of arrays is the group effect resulting from the combination of arrays multiplied by the directional characteristic of the individual arrays. The directional characteristic of the individual elementary

[1] The frequency range is limited at low frequencies by the fact that very little energy is coupled into the line because the collectors are so far out of resonance, and at high frequencies by the fact that, when the collectors approach resonance, the resulting reactive loading on the line alters the phase shift per unit distance along the transmission line, and hence makes it impossible to maintain the proper phasing.

array can be calculated as outlined in Sec. 133, while the group effect depends upon the arrangement of the elementary arrays and is the directional characteristic of an array composed of elements located at the center of each individual elementary array of the array of arrays, and having uniform radiation in all directions.

The commonest example of an array of arrays consists of two parallel broadside arrays one-quarter wave length apart and excited in phase quadrature. Such an arrangement is illustrated in Fig. 389*a* and can be considered as a broadside array, each element of which consists of two radiators spaced one-quarter of a wave length apart and connected so as to be one-quarter of a cycle out of phase. The directional characteristic in a horizontal plane of this two-element elementary antenna array is unidirectional, as shown in Fig. 389*b* and discussed in connection with

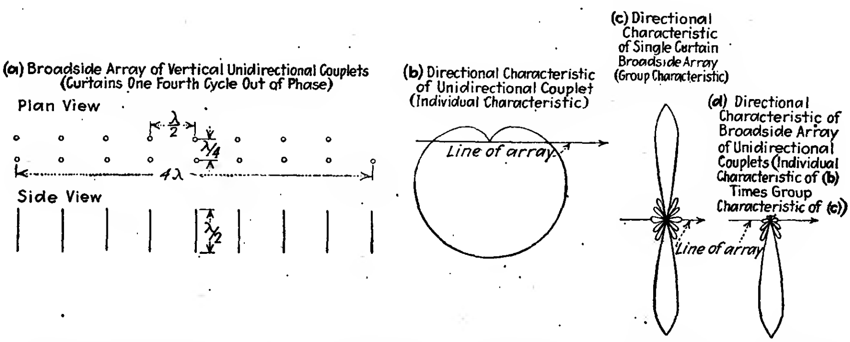


FIG. 389.—Broadside array, each element of which is composed of unidirectional couplets, together with polar diagrams showing individual and group characteristics, and the actual directional characteristic of the field radiated from the array of arrays.

Eq. (229), while the directional characteristic of a simple broadside array is as shown in Fig. 389*c*. The characteristic of the broadside array, each element of which consists of a unidirectional couplet, is obtained by multiplying the individual characteristic of Fig. 389*b* by the group directional characteristic of Fig. 389*c*, which results in the unidirectional broadside beam of Fig. 389*d*. The directional characteristic in a vertical plane is determined in a similar manner, but is not of great importance because the presence of the second broadside array does not materially affect the vertical radiation. Unidirectional characteristics in a horizontal plane, similar to those illustrated in Fig. 389, are obtained whenever the spacing between the two broadside arrays is an odd multiple of a quarter wave length and the phase difference between the two arrays is $\pm 90°$, depending on the desired direction.

Many short-wave antenna arrays consist of several arrays of broadside couplets stacked one above the other in order to give directivity in both vertical and horizontal planes. Such an antenna is illustrated in

Fig. 390 and can be considered as consisting of a broadside array of elementary antenna arrays, each element of which consists of an array of stacked unidirectional couplets. Each elementary array of such an antenna is in itself an array of arrays, and the directional characteristic is obtained by multiplying the directional characteristics of the unidirec-

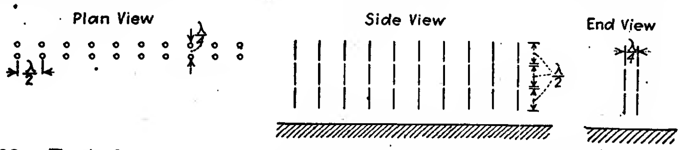


FIG. 390.—Typical antenna array consisting of a broadside array of stacked unidirectional couplets.

tional couplet of Fig. 389*b* by the directional characteristics of a stacked colinear antenna and then again multiplying by the directional characteristic of a simple broadside array. It will be observed that the principal effect of the stacking is to increase the directivity in the vertical plane, that the effect of the second curtain is largely confined to eliminating

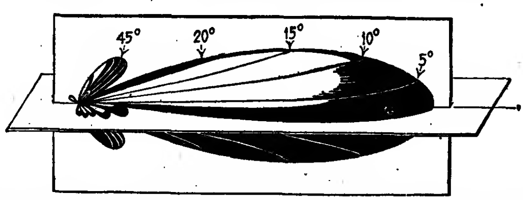


FIG. 391.—Directional characteristic in three-dimensional space of field radiated from a broadside array of stacked unidirectional couplets. (*From Southworth.*)

the back-end radiation, and that the broadside arrangement gives high directivity in a horizontal plane. The result is a very intense radiation confined to a small cone directed along the horizontal and aimed broadside to the antenna array. The directional characteristic of such an array in three-dimensional space is illustrated in Fig. 391

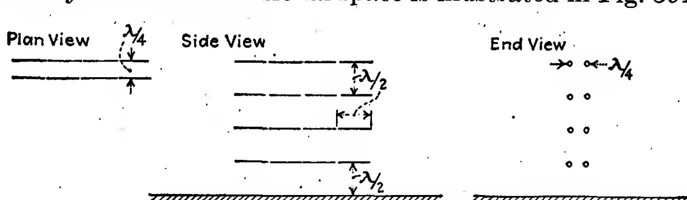


FIG. 392.—Views of a typical array composed of horizontally polarized antennas.

An array of arrays similar to that of Fig. 390, but consisting of horizontally polarized half-wave radiators is shown in Fig. 392. The resulting directional characteristics are much the same as in the case of vertical radiators, except that the ground reflection cancels the radiation at very low vertical angles.

The supporting towers of antenna arrays usually divide the main array into smaller arrays, or bays, between the ends of which there is a small spacing to accommodate the tower. Such an arrangement is essentially an array of arrays in which each bay represents an elementary array and the group effect corresponds to a linear array of elementary antennas with a spacing equal to the distance between bay centers. If the spacing between the adjacent bays is at all appreciable, parasitic lobes of considerable size appear in the directional characteristic and seriously reduce the gain of the array.[1]

The gain of an antenna array composed of an array of arrays depends upon the individual and group directional characteristics and follows a very complicated mathematical law. It is possible, however, to give certain general principles by which the gain of such antenna systems can be readily estimated under many practical conditions. Thus an array of undirectional couplets as in Fig. 389 has twice as much gain as the corresponding single-curtain broadside array. Again, when the elementary array tends to concentrate the radiation in one plane, while the group effect is to concentrate the radiation in another plane, the gain of the array of arrays is then very nearly equal to the gain of the elementary array multiplied by the gain of the group array. Thus, when several broadside arrays of couplets are stacked, the total gain is very nearly the gain obtained by stacking elements in a colinear array, multiplied by 2 to take into account the effect of the second broadside array, and finally multiplied by the gain of a linear broadside array corresponding to the group array.

Antennas Employing Non-resonant Wires; Rhombic Antennas.[2]— A number of directive antenna systems have been developed based upon non-resonant radiating elements such as discussed in connection with Fig. 381. These arrangements have the advantage of operating over a considerable frequency band without readjustment, and so are particularly desirable under conditions where different day and night frequencies must be used.

[1] See G. C. Southworth, Certain Factors Affecting the Gain of Directive Antennas, *Proc. I.R.E.*, vol. 18, p. 1502, September, 1930.

[2] Further information on rhombic and non-resonant V antennas is given by E. Bruce, Developments in Short-wave Directive Antennas, *Proc. I.R.E.*, vol. 19, p. 1406, August, 1931; E. Bruce, A. C. Beck, and L. R. Lowry, Horizontal Rhombic Antennas, *Proc. I.R.E.*, vol. 23, p. 24, January, 1935.

In reading these papers it is to be kept in mind that the method of analysis presented involves the approximation that the current is uniformly distributed along the wire. This is equivalent to assuming that there is no radiation from the antenna, an obvious absurdity. However, with practical proportions of rhombic antennas, the directional characteristics calculated on this basis do not differ greatly from the true directional characteristic, so the approximation is permissible for this case.

A simple arrangement of this sort is illustrated in Fig. 393*a* and consists of two wires arranged in the form of a V and made non-resonant by grounding the far ends through resistances equal to the characteristic impedance of the wires. By suitably tilting the wires the radiations from the two legs will add in the direction of the bisector of the apex angle, as shown in Fig. 393, while in other directions the radiation will tend to cancel. This optimum value of the tilt angle Φ is given by the solid line in Fig. 394 as a function of the wire length.[1] When the length of the leg exceeds about two wave lengths, the angle is seen to vary only slowly with the length measured in wave lengths. As a consequence a

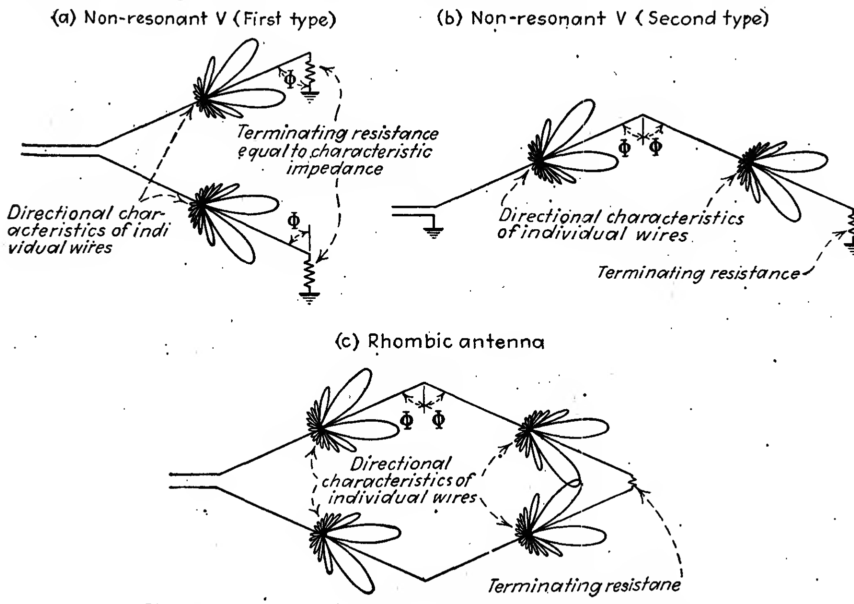


FIG. 393.—Various forms of non-resonant antenna systems.

non-resonant V with long legs will maintain its directive pattern reasonably well over a considerable frequency range.

A modification of the non-resonant V antenna is shown in Fig. 393*b* and consists in exciting the system from one end of the V and terminating the other end through the characteristic impedance. In this arrangement the major lobes are in the same direction when the tilt angle Φ has

[1] This figure gives the optimum tilt angle for antennas in *free space*. When the plane of the antenna is parallel to the surface of the earth and the antenna is near earth (which is the usual case), the optimum tilt angle is smaller than given in Fig. 394 by an amount that depends on the height.

These tilt angles as given in Fig. 394 do not contain the approximation which Bruce states was made in deriving the curves of optimum tilt given in his 1931 paper, and so are more accurate.

the value given by the solid line in Fig. 394, while the major lobes of the two legs add up in phase when the tilt angle has the value given by the dotted curve in Fig. 394. The optimum tilt angle is hence somewhat intermediate between these two curves. The directivity with this arrangement is not particularly sensitive to frequency (or tilt angle) because changes produced in the direction of the major lobe of one leg tend to be counteracted by changes of opposite character in the direction of the major lobe of the other leg.

The most widely used form of non-resonant antenna is illustrated in Fig. 393*c* and is made up of four non-resonant wires arranged in the form of a diamond or *rhomboid*. This arrangement can be considered as being built up of two non-resonant V's of Fig. 393*b* placed side by side. The rhomboid arrangement gives increased directivity and simplifies the terminating problem by avoiding the necessity of providing a ground connection that will be independent of frequency and have the same resistance in all weather. The directional characteristic of the radiation from each leg of the rhomboid is as shown, and, when the tilt angle Φ has the optimum value according to Fig. 394, all four legs have major lobes in the direction of a line drawn through the apexes, as shown, and tend to add up in phase.

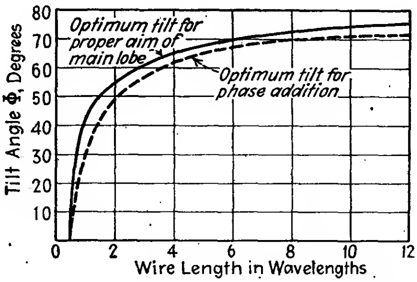


FIG. 394.—Optimum tilt angle for non-resonant antennas in space (assuming zero attenuation).

The directional characteristics of a rhomboid antenna are not particularly critical with respect to tilt angle provided the length of each leg is not less than two wave lengths, and as a consequence the rhomboid can be used over a considerable frequency range without any adjustment. The effect of a change in frequency is to alter the length of the legs when measured in wave lengths. This alters the angle that the lobes of maximum radiation make with respect to their corresponding wires, but, as is apparent from Fig. 393*c*, the two legs that are on the same side of the diamond are affected oppositely and also no longer add exactly in phase. However, since the phase difference and the change in tilt angle with frequency are small when the legs are long, the result of a change in frequency is merely to alter the sharpness of the directional pattern somewhat without changing its fundamental character. The frequency range obtainable is approximately 2 to 1.

Typical directional characteristics obtained with the same rhombic antenna operated at different frequencies are shown at Fig. 395. The greatest directivity is obtained at the highest frequency, where the length

of the legs measured in wave lengths is greatest, but the general character of the directional pattern is maintained without change over the 2-to-1 variation in frequency. In the usual case where the plane of the rhombic antenna is parallel to the earth, the radiation is horizontally polarized and the most important factor controlling the directivity in a vertical plane is the height above ground. The tilt angle and the length of the legs also influence the vertical directivity, however, and can be so chosen as to compensate to a considerable extent for lack of proper height.[1]

The terminating resistor for the rhombic and other non-resonant antennas must equal the characteristic impedance of the line if the ideal unidirectional characteristic is to be obtained. It is possible, however, to modify the minor lobes to the rear, and in particular to obtain a null in any desired backward direction, by merely modifying the magnitude or phase angle, or both, of the terminating resistance a slight amount as experiment indicates is required. When the rhombic antenna is used with a transmitter, the terminating resistance must

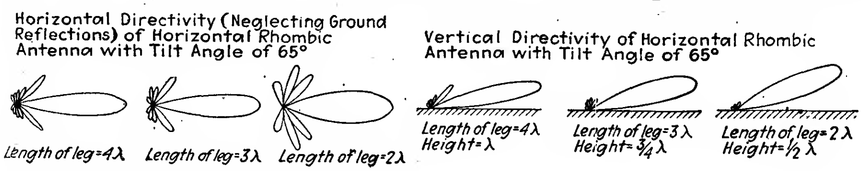


FIG. 395.—Polar diagrams showing directional characteristics of the same horizontal diamond-shaped antenna for three different frequencies.

absorb all the power that would otherwise be radiated in the backward direction. The amount of energy involved depends upon the length of the legs, but in the usual designs is of the order of half the total power supplied.

In practical rhombic antennas it is desirable to make up each conductor of two wires connected in parallel and spaced a small distance apart. The characteristic impedance of the antenna can then be controlled by the spacing, which may be made such as to assist in matching the radio-frequency transmission line. Furthermore with an ordinary wire the characteristic impedance at different positions along the leg is not the same because of the varying spacing between opposite legs as one moves along the diamond. With the spaced-wire conductor it is possible to compensate for this effect by varying the spacing between parallel conductors in such a way as to keep the characteristic impedance constant. This means a greater spacing as the wire recedes from ground, or, in the case of the diamond, as the two sides separate.

The directional characteristics of rhombic and other non-resonant antenna systems can be calculated by the same principles that have

[1] See Bruce, *loc. cit.*

been applied to other types of antennas. Formulas for the rhombic antenna are found in Sec. 133.

The gain of a rhombic antenna depends upon the design and the length of the legs, and in commercial designs is about 30 to 40 compared with a half-wave radiator.

The V (or Folded Wire) Antenna.[1]—This type of antenna is illustrated in Fig. 396 and consists of two long resonant wires arranged to form a V and excited so as to carry equal currents that are in phase opposition. The radiation from the individual wires of the V antenna has the character illustrated in Fig. 375, and it is seen that with a suitable choice of apex angle there is a concentration of radiation in the direction of the bisector of the apex angle.[2] The resulting directional pattern is as illustrated,

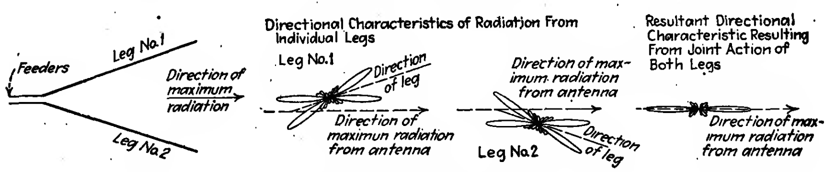


FIG. 396.—Resonant V antenna, showing how the radiation from the two legs combines to give a well-defined beam.

and possesses very marked directivity in the plane of the V, particularly when the legs are long. Some directivity is also obtained in the vertical plane although this is less pronounced.

Increased directivity can be obtained by means of an array, each element of which is a V antenna. Thus the backward radiation can be eliminated by the use of two V's spaced an odd number of quarter wave lengths apart and excited 90° out of phase to give an end-fire action (Fig. 397).[3] The directivity in a vertical plane can be improved by stacking two or more V's above each other as illustrated in Fig. 397*a*. The

[1] See P. S. Carter, C. W. Hansell, and N. E. Lindenblad, Development of Directive Transmitting Antennas by RCA Communications, Inc., *Proc. I.R.E.*, vol. 19, p. 1773, October, 1931; P. S. Carter, Circuit Relations in Radiating Systems, *Proc. I.R.E.*, vol. 20, p. 1004, June, 1932.

[2] The apex angle is ordinarily made equal to twice the angle that exists between the major lobe of radiation from one leg of the V, considered as an isolated radiator, and the axis of the wire. This is the optimum apex angle for a V antenna in free space when the legs are at least 2λ long. In the practical case when the antenna is near the earth with the plane of the V parallel to the earth's surface, the ground reflections introduce modifications that tend to reduce the optimum angle somewhat.

[3] An alternative arrangement for eliminating back-end radiation is shown in Fig. 397*d*, and consists in placing a second V a short distance above or below the first. This second V is a half wave length shorter than the first V, and is excited in phase quadrature. Since its center is almost exactly a quarter wave length closer to the apex than the center of the other V, the two acting together are essentially equivalent to the arrangement of Fig. 397*a*, but only half as many supporting poles are required.

vertical directivity is also determined to a considerable extent by the height above earth, since the radiation from a horizontal V is horizontally polarized. Greater directivity in a horizontal plane can likewise be realized by arranging V's in broadside as illustrated in Fig. 397*b*.

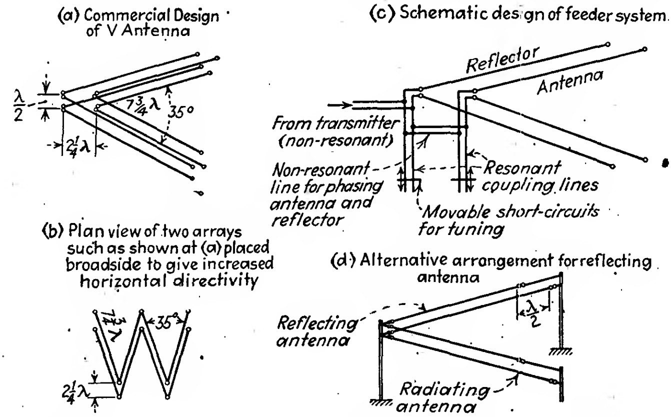


FIG. 397.—Commercial designs of resonant V antenna array, together with feeder circuits.

The directional characteristics of the V antenna can be calculated by adding the radiated fields produced by the individual legs, being careful to take into account the fact that the legs are not parallel. Formulas are given in Sec. 133.

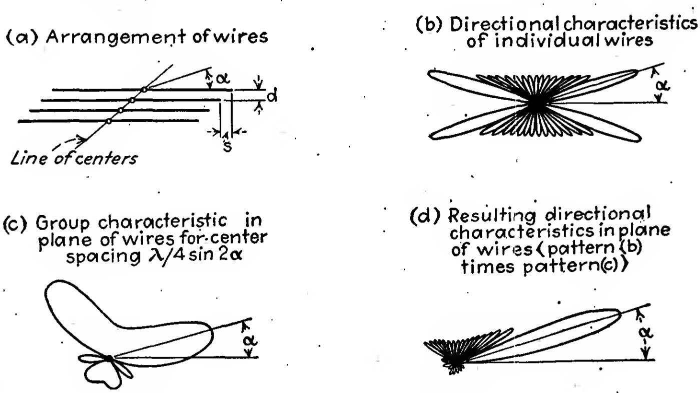


FIG. 398.—Harmonic-wire array, with field patterns illustrating how the directional pattern is built up.

The gain of a bay such as illustrated in Fig. 397*a* located at an average height of one-half wave length is approximately 39 times as compared with a half-wave antenna.

Harmonic-wire Arrays.[1]—This array is an ingenious combination of

[1] See Carter, Hansell, Lindenblad, *loc. cit.*

four long wires arranged in parallel in a staggered formation as shown in Fig. 398*a* and carrying equal currents with a progressive phase difference of 90° between wire currents. The wires are arranged so that the maximum of radiation from an individual wire when isolated is in the desired direction as indicated, while the spacing between wire centers is[1]

$$\text{Distance between wire centers} = \frac{\lambda}{4 \sin 2\alpha} \tag{234}$$

where α is the angle of maximum radiation from single wire with respect to wire axis as given by Fig. 376*a*. The resulting directional characteristic is obtained by multiplying the directional characteristic of the individual wire as shown at Fig. 398*b* by the directional characteristic of four non-directional radiators spaced $\frac{\lambda}{4 \sin 2\alpha}$ apart with a progressive phase difference of 90°, corresponding to the arrangement of the actual antennas. This latter characteristic is shown at (*c*) for the plane of the wires, and the resulting directional characteristic of the harmonic-wire array in the plane of the wires is as at (*d*). It is to be noted that the unidirectional characteristic of (*d*) comes from the fact that, with the phasing and spacing specified, the group characteristic always has a zero in the backward direction. There is no radiation at right angles to the plane of the wire because of the way the wires are spaced.

Commercial designs of harmonic-wire arrays are shown in Fig. 399. A single bay composed of four wires approximately eight wave lengths long has a gain of approximately 16 as compared with a half-wave comparison antenna.

Polyphase Antenna Arrays.—Polyphase antenna systems have been suggested for meeting certain special requirements. Thus the *turn stile* array of Fig. 400 has been proposed for ultra-high-frequency broadcast transmission, where very high vertical directivity is desirable with substantially uniform radiation in all horizontal directions.[2] This arrangement consists of two arrays oriented 90° in space and excited 90° out of phase (*i.e.*, two-phase system). Each array consists of stacked horizontal half-wave antennas carrying identical currents, and thus gives a sharp directivity in the vertical plane. The horizontal directivity of the individual arrays is as shown in Fig. 400, but, since these combine in phase quadrature, the combined directional pattern is very nearly circular.

[1] The stagger distance s and wire spacing d that correspond are $\frac{\lambda \sin \alpha}{4 \sin 2\alpha}$ and $\frac{\lambda \cos \alpha}{4 \cos 2\alpha}$, respectively.

[2] See George H. Brown, The Turnstile Antenna, *Electronics*, vol. 9, p. 15, April, 1936.

Other polyphase arrangements can be devised to meet different requirements. Thus at broadcast and lower frequencies it is possible to obtain a large amount of vertical directivity by a polyphase array of short radiators such as shown in Fig. 401.[1] Here there are a number of short vertical radiators associated in a polyphase manner to obtain a substantially non-directional characteristic in the horizontal plane while maintaining marked directivity in the vertical. Directional character-

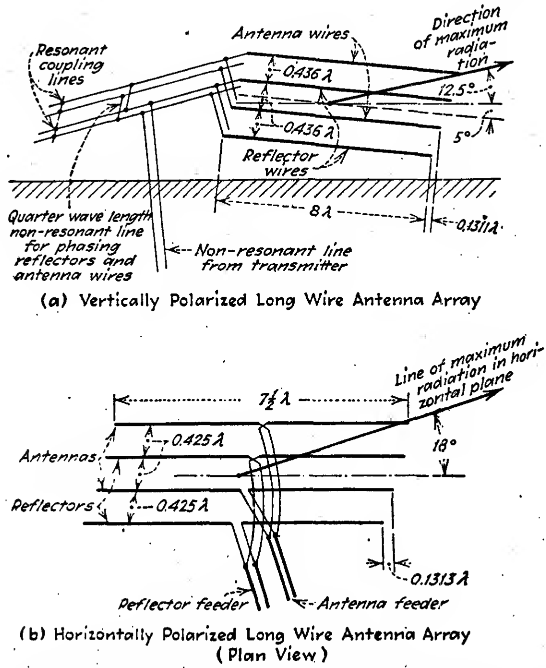


FIG. 399.—Examples of harmonic-wire antenna arrays, together with feeder circuits for the vertically polarized array. (Feeder arrangement for horizontal array is similar.) The spacings shown between wires represent distances between corresponding points, and so are measured at a stagger from the perpendicular.

istics for a typical case are shown at Fig. 401. Polyphase arrangements of this type are very flexible, since by simple changes in the array it is possible to vary the directional characteristics in almost any manner desired.

Beverage or Wave Antenna.[2]—This was the first non-resonant antenna, and it is used in the reception of long-wave signals. The Beverage or

[1] Arrangements of this type were suggested by the author in a paper presented before the Seattle section of the Institute of Radio Engineers, Mar. 27, 1936.

[2] An extensive discussion of the wave antenna, including a complete mathematical analysis, practical construction data, etc., is given by Harold H. Beverage, Chester W. Rice, and Edward W. Kellogg, The Wave Antenna, *Trans. A.I.E.E.*, vol. 42, p. 215, 1923.

wave antenna consists of a wire ranging from one-half to several wave lengths long, pointed in the direction of the desired transmitting station and mounted at a convenient height (usually 10 to 20 ft.) above earth. The end toward the transmitting station is grounded through an impedance approximating the characteristic impedance of the antenna when

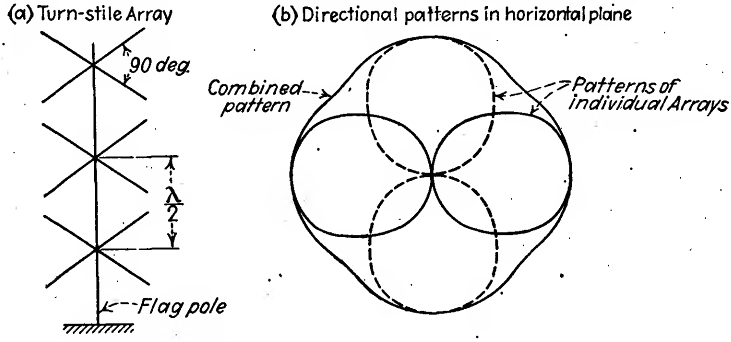


FIG. 400.—Turnstile array (two-phase system) together with directional patterns in horizontal plane.

considered as a one-wire transmission line with ground return, while the energy abstracted from the radio waves is delivered to a radio receiver at the end of the antenna farthest from the transmitter.

The wave antenna abstracts energy from passing waves as a result of the wave-front tilt which the earth losses produce in vertically polarized

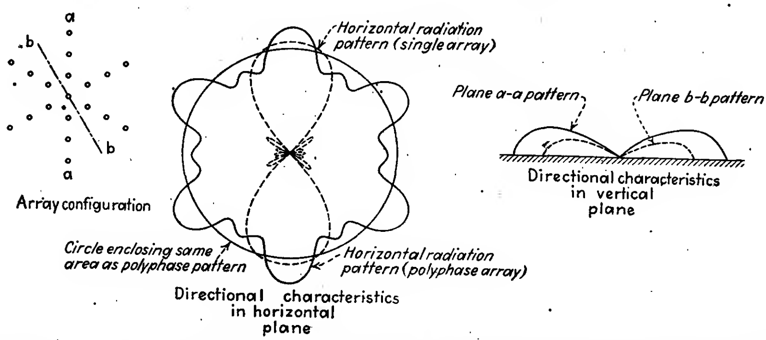


FIG. 401.—Polyphase array which concentrates radiation along the horizontal without requiring high pole structures. By varying the angle between the lines of antennas, or by modifying the arrangement of antennas within a line, etc., it is possible to vary the horizontal and vertical patterns as desired.

waves traveling along the surface of the ground. This tilt causes voltages and hence currents to be induced in the wire. When the wave is traveling in the direction of the wire toward the receiver as in Fig. 402, the induced currents all add up in phase at the receiver because the currents induced in the wire travel with the same velocity as the radio wave and so all

keep in step with it. When the wave travels in the opposite direction, the energy builds up as the terminated end is approached, but the terminating resistance absorbs this energy so that there is little or no effect on the receiver. Waves arriving from the side induce currents that largely cancel out at the receiver because, although the various induced currents are more or less in phase at their point of origin, they are at widely different distances from the receiver.

The wave antenna has marked directivity in the horizontal plane, as shown in Fig. 402, and this can be still further improved by arranging several wave antennas in a simple array.[1] Since the wave antenna is non-resonant, it can be used to receive simultaneously signals of different wave lengths provided that the transmitters all lie in the same general direction.

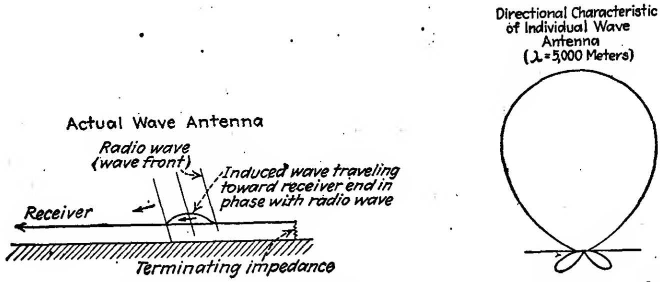


FIG. 402.—Details of action taking place in the Beverage wave antenna when the incident wave arrives from the direction in which the response is maximum, together with directional diagram when wire length is 0.9 wave lengths.

133. Methods and Formulas for Calculating the Directional Characteristics of Antenna Systems.—The general method of calculating the directional characteristics of antenna systems has already been outlined in Sec. 129, and applied to specific cases in the two preceding sections. Briefly it consists in determining the magnitude and phase of the field produced at a distant point by each elementary length of the antenna, and then adding the resulting fields vectorially for all the elementary lengths of the entire antenna system. This method can be applied irrespective of the geometry involved, the number of antennas, the type of antennas, the current distribution, etc.

Effect of Ground.—When an antenna is near the ground, waves radiated toward the earth are reflected and combine with the unreflected wave as discussed in connection with Fig. 372. In the case of transmitting antennas it is convenient to consider the reflected wave as being produced by an image antenna as discussed in Sec. 128 and illustrated in

[1] See Austin Bailey, S. W. Dean, and W. T. Wintringham, The Receiving System for Long-wave Trans-atlantic Radio Telephony, *Proc. I.R.E.*, vol. 16, p. 1645, December, 1928.

Figs. 372 and 373.[1] In the case of a perfect ground the currents in corresponding parts of the actual and image antennas are the same, but have opposite phase when these parts are horizontal and the same phase when the parts are vertical, as indicated in Fig. 373. It will be noted that with ungrounded wires the antenna plus its image forms a simple array consisting of two identical antennas with centers spaced a distance equal to twice the height of the actual antenna above earth. With horizontal antennas and also vertical antennas that are an even number of half wave lengths long, the antennas can be considered as being excited 180° out of phase (negative image antenna), while with vertical antennas that are an odd number of half wave lengths long the two antennas can be considered as being excited in the same phase (positive image). The effect of a perfect ground on the directional characteristic of an ungrounded antenna can hence be obtained by multiplying the free-space characteristic of the antenna by the group characteristic corresponding to an array having two suitably spaced elements radiating uniformly in all directions. This gives[2]

For negative image:

$$\left.\begin{matrix}\text{Actual radiation in}\\ \text{presence of ground}\end{matrix}\right\} = 2 \sin\left(2\pi\frac{H}{\lambda}\sin\theta\right)\left\{\begin{matrix}\text{Radiation from}\\ \text{antenna when}\\ \text{in free space}\end{matrix}\right. \qquad (235a)$$

For positive image:

$$\left.\begin{matrix}\text{Actual radiation in}\\ \text{presence of ground}\end{matrix}\right\} = 2 \cos\left(2\pi\frac{H}{\lambda}\sin\theta\right)\left\{\begin{matrix}\text{Radiation from}\\ \text{antenna when}\\ \text{in free space}\end{matrix}\right. \qquad (235b)$$

where H/λ is the height of the antenna center above ground, measured in wave lengths, and θ is the angle of elevation above the horizontal. The nature of the factor $\left|2 \sin\left(2\pi\frac{H}{\lambda}\sin\theta\right)\right|$, which takes into account the effect of a negative image, is indicated in Figs. 403 and 404. It is seen that the ground reflection causes cancellation along the ground and at certain vertical angles. The angle of elevation of the first lobe above the horizontal decreases as the height of the antenna above earth

[1] Directional characteristics obtained in this way on the basis of a radiating antenna are, according to the Reciprocity Theorem, the same as the directional characteristics obtained when the same antenna is used for reception. In the case of receiving antennas the wave that is effective in inducing voltage in the antenna is the vector sum of a direct and a reflected wave, with the latter being related to the direct wave through the reflection coefficients of the earth.

[2] These equations follow at once from Eq. (229) by redefining Φ, noting that the spacing $d = 2H$, and assuming $\alpha = 180°$ for a negative image and 0° for a positive image.

is increased. Hence, to obtain strong radiation in directions approaching the horizontal in the presence of a negative image, it is necessary that the height above earth be at least one wave length. With a positive image the positions of the lobes and nulls are interchanged from the conditions shown in Figs. 403 and 404, and there is a maximum of radiation along the ground. When the antenna is grounded, short-cut methods such as illustrated by Eqs. (235*a*) and (235*b*) cannot be used, and the total field is obtained by summing up the radiation from the actual antenna plus the radiation from the image.

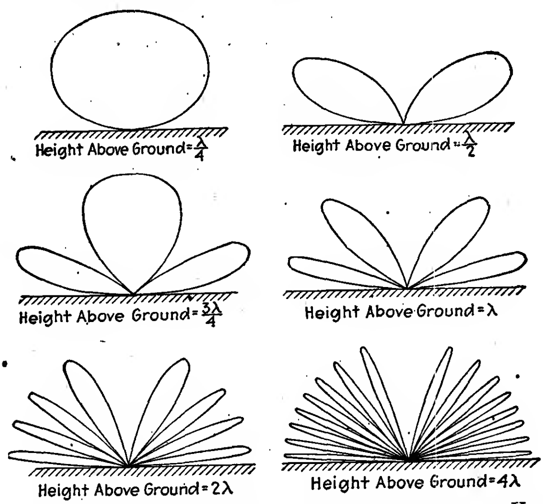


FIG. 403.—Polar diagram of the negative image factor $|2 \sin (2\pi \frac{H}{\lambda} \sin \theta)|$ for various values of $\frac{H}{\lambda}$ showing how the height above earth affects the directional characteristic of the antenna.

In the practical case of an imperfect earth, losses in the ground cause the reflected wave to be smaller in magnitude and slightly different in phase from the reflected wave obtained with a perfectly conducting earth. With the aid of the Reciprocity Theorem it can be shown that the reflection coefficients of Eq. (216) hold except in directions where a Sommerfeld type of ground wave exists, even though these equations were derived on the basis of plane waves while the waves near a transmitting antenna are obviously not plane.[1] As a result of the incomplete

[1] The reasoning leading to this conclusion follows: When the antenna receives waves from a distant transmitter, the resulting field at the antenna is the sum of a direct wave and a reflected wave, and the latter can be determined with the aid of the reflection coefficients of Eq. (216) because the waves are unquestionably plane. However, according to the Reciprocity Theorem the behavior is the same when the antenna is transmitting as when receiving, so it follows that the reflection coefficients must be the same when radiating as when receiving. This analysis fails to hold

reflection, the currents that can be considered as flowing in the image antenna are less than the image currents with perfect reflection by a factor A, and likewise differ in phase from the perfect ground case by an angle Φ, where A and Φ are the magnitude and phase of the reflection

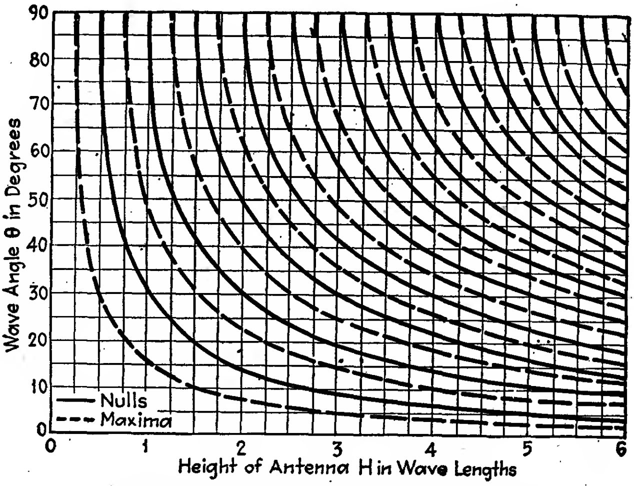


FIG. 404.—Chart showing the vertical angles at which the negative image factor $|2 \sin (2\pi \frac{H}{\lambda} \sin \theta)|$ is maximum and zero. These represent the vertical angles at which the reflections from the ground cause complete reinforcement and complete cancellation, respectively, of the radiation. With a positive image the position of the nulls and maxima are interchanged.

coefficients as calculated in Eq. (216). The equation giving the resultant sum of the direct and reflected waves under such conditions is:[1]

$$\left.\begin{array}{l}\text{Actual radiation}\\ \text{in presence of}\\ \text{ground}\end{array}\right\} = \left[1 + A^2 \pm 2A \cos\left(\Phi + 4\pi\frac{H}{\lambda}\sin\theta\right)\right]^{1/2} \left\{\begin{array}{l}\text{Radiation}\\ \text{from}\\ \text{antenna}\\ \text{when in}\\ \text{free space}\end{array}\right. \tag{236}$$

when a ground wave is formed, *i.e.*, with vertically polarized waves at low vertical angles, because the Sommerfeld type of ground wave is not a plane wave even when produced by a distant transmitter, and so does not follow the ordinary optical laws upon which the reflection coefficients of Eq. (216) are based.

[1] This equation is obtained by noting that the resultant wave is the sum of two components having relative amplitudes of unity and A, with a phase difference between them of $\left(\Phi + 4\pi\frac{H}{\lambda}\sin\theta\right)$ for a positive image and $\left(180° + \Phi + 4\pi\frac{H}{\lambda}\sin\theta\right)$ for a negative image. The addition of these two vectors by the formula giving the third side of a triangle when two sides and the included angle are known leads at once to Eq. (236)

The + sign is for a positive image and the negative sign for a negative image, while the remaining notation is the same as used above.

In making use of Eq. (236) it is to be noted that the magnitude and phase of the reflection coefficient depend upon the angle of incidence at the reflecting surface and upon the polarization of the wave. Hence the currents that are assumed to flow in the image antenna vary with the direction in which the resultant radiation is desired, and under some circumstances the situation becomes quite complex. Thus with a horizontal wire the radiation is horizontally polarized in a vertical plane perpendicular to the axis of the wire, is vertically polarized in a vertical plane containing the wire, and has both vertical and horizontal components in intermediate vertical planes. In the latter case the reflection is calculated separately for vertically and horizontally polarized components, and then is combined with the direct wave to obtain the total resultant fields.

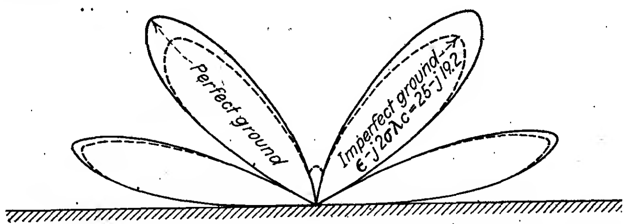


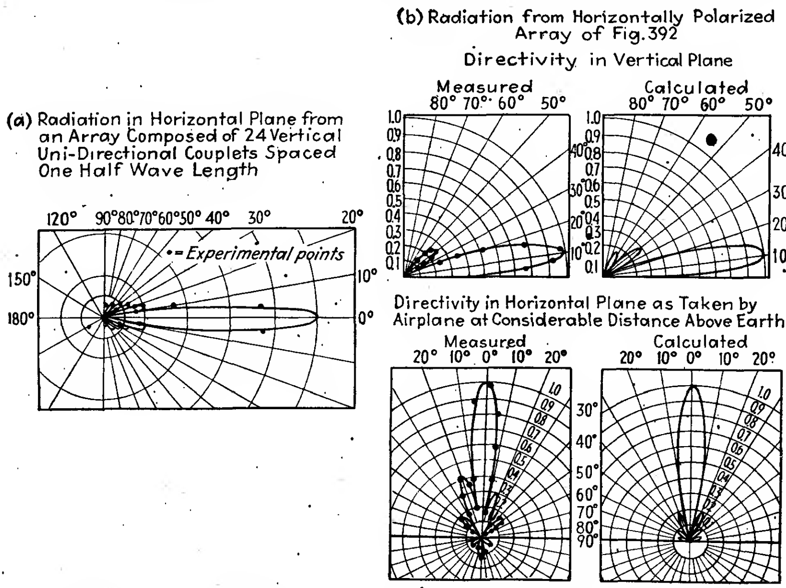
FIG. 405.—Curves showing the effect of ground losses upon the negative image ground factor with a horizontally polarized antenna one wave length above a typical earth.

When a ground wave is formed, as is the case when vertically polarized waves are radiated at or near the horizontal, the reflection-coefficient method of analysis fails to hold, and it is necessary to employ the Sommerfeld analysis. The intensity of the ground wave along the horizontal in the immediate vicinity of the antenna is then the same for an imperfect earth as for a perfect earth, while at higher angles of elevation the ground wave gradually merges into the wave calculated on the basis of reflection coefficients, until at vertical angles appreciably above the pseudo-Brewster angle it is possible to ignore the ground wave. It will be noted that ground waves are produced only when vertically polarized waves are radiated along the horizontal, and that with horizontally polarized waves the reflection-coefficient method of analysis applies even at very low vertical angles.

The principal effect of the ground losses in practical cases is to reduce the field intensities slightly while leaving the shape of the directional characteristic substantially unchanged. This is illustrated in Fig. 405, which compares the fields obtained for a perfect earth with those existing with typical earth. The principal effects produced by

the ground losses are seen to be a reduction in the amplitude of the lobes, together with a slight filling in of the nulls. The discrepancy approaches zero at low vertical angles, and is so small even at high vertical angles as not to alter appreciably the shape of the directional pattern. As a consequence, the directional patterns of antennas are usually calculated on the assumption of a perfect earth.

Agreement between Calculated and Observed Antenna Characteristics.—Experience has shown that the gain and the directional characteristics of antenna arrays calculated on the basis of perfectly conducting earth



406.—Comparison of measured and calculated distribution of radiated field about arrays composed of horizontal and vertical elements. The calculated results assume a perfect earth.

agree satisfactorily with observed results under nearly all conditions. Comparisons of theoretical and observed directional characteristics for typical vertically and horizontally polarized antenna arrays[1] are given in Fig. 406 and show excellent agreement.

Formulas for Calculating Radiated Field and Directional Characteristics of Common Antenna Systems.—1. Radiation from an elementary length of wire in the absence of ground [from Eq. (221)]:

$$\epsilon = \frac{60\pi}{d\lambda} I(\delta l) \cos \theta \tag{237}$$

[1] See Southworth, *loc. cit.;* see also M. Bäumler, K. Kruger, H. Plendl, and W. Pfitzer, Radiation Measurements of a Short-wave Directive Antenna at the Nauen High-power Radio Station, *Proc. I.R.E.*, vol. 19, p. 812, May, 1931.

where

ϵ = field strength in volts per meter
d = distance to antenna in meters
θ = angle of elevation with respect to plane perpendicular to antenna wire
λ = wave length of radiated wave in meters
δl = length of elementary antenna in meters
I = current in antenna in amperes.

2. Radiation from a wire in space.[1]

a. When the wire is an odd number of half wave lengths long:

$$\epsilon = \frac{60I}{d} \frac{\cos\left(\pi\frac{L}{\lambda}\cos\theta\right)}{\sin\theta} \tag{238a}$$

b. When the wire is an even number of half wave lengths long [from Eq. (228)]:

$$\epsilon = \frac{60I}{d} \frac{\sin\left(\pi\frac{L}{\lambda}\cos\theta\right)}{\sin\theta} \tag{238b}$$

where

ϵ = field strength in volts per meter
d = distance to antenna in meters
I = current in amperes at a current loop
L = length of antenna in meters
λ = wave length in meters
θ = angle of elevation measured with respect to wire axis.

3. Radiation from vertical grounded wire, assuming a perfect earth:[2]

$$\epsilon = \frac{60I}{d}\left[\frac{\cos 2\pi\frac{L}{\lambda} - \cos\left(2\pi\frac{L}{\lambda}\sin\theta\right)}{\cos\theta}\right] \tag{239}$$

where θ is the angle of elevation with respect to ground and the remainder of the notation is as in Eq. (237).

4. Radiation from a non-resonant wire in space [from Eq. (233)]:

$$\epsilon = \frac{30I_0}{d\sqrt{1 + (\alpha/p)^2}} \frac{\sin\theta}{(1 - \cos\theta)}(1 + \epsilon^{-2\alpha L} - 2\epsilon^{-\alpha L}\cos pL)^{1/2} \tag{240a}$$

[1] See Carter, Hansell, and Lindenblad, *loc. cit.*

[2] See Stuart Ballantine, On the Radiation Resistance of a Simple Vertical Antenna at Wave Lengths Below the Fundamental, *Proc. I.R.E.*, vol. 12, p. 823, December, 1924.

where

ϵ = field strength in volts per meter
I_0 = current entering the line
d = distance to antenna in meters
α = attenuation constant of line in meters
p = $(2\pi/\lambda)(1 - \cos\theta)$
θ = angle with respect to wire axis
L = length of wire in meters
λ = wave length in meters (assumed same on wire as in space).

In the special case where the attenuation α can be assumed zero (*i.e.*, when the wires are so short that the current can be considered to be constant),

$$\epsilon = \frac{60 I_0}{d} \frac{\sin\theta}{1 - \cos\theta} \sin\left[\pi\frac{L}{\lambda}(1 - \cos\theta)\right] \qquad (240b)$$

5. Radiation from a rhombic (diamond) antenna in space, with characteristic impedance termination, and neglecting the attenuation along the wires:[1]

$$\left.\begin{array}{l}\text{Relative field}\\\text{strength}\end{array}\right\} = \left(\frac{\cos(\Phi - \beta)}{1 - \cos\theta\sin(\Phi - \beta)} + \frac{\cos(\Phi + \beta)}{1 - \cos\theta\sin(\Phi + \beta)}\right)$$
$$\times \sin\left\{\frac{\pi l}{\lambda}[1 - \cos\theta\sin(\Phi + \beta)]\right\} \sin\left\{\frac{\pi l}{\lambda}[1 - \cos\theta\sin(\Phi - \beta)]\right\} \qquad (241)$$

where

Φ = tilt angle of antenna (see Fig. 393*c*)
β = bearing angle with respect to line passing through the apex having the terminating resistance and the diagonally opposite apex
θ = angle of elevation with respect to the plane of the antenna
l = length of each side of the rhomboid
λ = wave length.

6. Radiation from a resonant V antenna (see Fig. 396) when the length of each leg is an even multiple of a half wave length and the antenna is remote from ground:[2]

Field strength in plane of V =

$$\sqrt{E_a^2 + E_b^2 - 2E_aE_b\cos\left(2\pi\frac{l}{\lambda}\sin\alpha\sin\Phi\right)} \qquad (242a)$$

[1] See Bruce, Beck, and Lowry, *loc. cit.*
[2] See Carter, Hansell, and Lindenblad, *loc. cit.*

where

E_a and E_b = radiation in desired direction from individual legs of antenna as given by Eq. (238)

l = length of leg

λ = length corresponding to one wave length

α = half of angle at apex

Φ = bearing angle with respect to bisector of apex.

Radiation in vertical plane passing through bisector of apex angle =

$$\frac{120I}{d}\left[\frac{\sin\left(\frac{n\pi}{2}\cos\alpha\cos\theta\right)\sin\alpha}{1-\cos^2\theta\cos^2\alpha}\right] \tag{242b}$$

where

n = number of half wave lengths in each leg of antenna

α = half of angle of apex

θ = angle of elevation with respect to plane of antenna

I = current at current loop

d = distance to antenna in meters.

7. Radiation from an antenna array remote from ground and consisting of $\mathfrak{N}$ parallel planes each made up of N parallel columns where each column is made up of n individual radiating elements radiating uniformly in all directions with equal intensity:[1]

$$\text{Relative field strength} = \left|\frac{\sin n\pi(a\cos\Phi\cos\theta+b)}{n\sin\pi(a\cos\Phi\cos\theta+b)}\right.$$
$$\left.\frac{\sin N\pi(A\sin\Phi\cos\theta+B)}{N\sin\pi(A\sin\Phi\cos\theta+B)}\,\frac{\sin\mathfrak{N}\pi(\mathfrak{A}\sin\theta+\mathfrak{B})}{\mathfrak{N}\sin\pi(\mathfrak{A}\sin\theta+\mathfrak{B})}\right| \tag{243}$$

where

n, N, and $\mathfrak{N}$ = number of radiators along x-, y-, and z-axes, respectively

a, A, and $\mathfrak{A}$ = spacing of adjacent radiators along x-, y-, and z-axes, respectively, measured in fractions of a wave length

b, B, and $\mathfrak{B}$ = phase displacement between adjacent radiators along x-, y-, and z-axes, respectively, measured in fractions of a cycle

θ = angle with respect to xy-plane (angle of elevation)

Φ = angle with respect to xz-plane (bearing angle).

8. Radiation from an antenna array in which the individual antennas are not spherical radiators is obtained by calculating the directional characteristic for spherical radiators, using Eq. (243) and then multiplying the result by the actual directional characteristic of the individual antenna involved.

[1] See Southworth, *loc. cit.*

9. Radiation from an array of arrays is determined by calculating the individual directional characteristic of the elementary array as outlined in Eq. (243) above, and then multiplying this individual effect by the group effect calculated by determining the directional characteristic of an array consisting of spherical radiators located at the centers of the elementary arrays.

10. Loop antennas are considered in Sec. 139.

11. The effect of the ground in the case of ungrounded antennas can be taken into account by the methods discussed above in connection with Eqs. (235) and (236). The ground effect is already included in Eq. (239) for the grounded antenna.

134. Power Relations in Transmitting Antennas.—The power relations existing in an antenna can be derived either from the radiated fields or from the self and mutual impedances existing in the antenna system.[1]

Evaluation of Radiated Power, Radiation Resistance, and Gain in Terms of the Radiated Field.—The procedure for carrying out this method has already been indicated in Sec. 129 and is briefly as follows: The antenna is assumed to be at the center of a large sphere, a convenient antenna current is assumed, and the resulting fields produced over the spherical surface are calculated. The energy passing through each square centimeter of the spherical surface is then $0.00265\epsilon^2$ watts, where ϵ is the field strength in r.m.s. volts per centimeter. The total energy radiated through the entire spherical surface is then found by a process of summation. This summation can be carried out either by mathematical or by graphical methods. In the mathematical method the appropriate expression for field strength as given in Sec. 133 is squared and integrated over the surface of the sphere.[2] This is a straightforward process, but even in the simplest practical cases it involves seldom used mathematical functions, and in the more complicated cases the indicated integrations cannot be performed with known functions.

Graphical means of summation can be used when mathematical methods are impossible or impracticable, and they are relatively simple if a systematic procedure is followed. The *first step* is to determine the average effective field strength at the spherical surface as a function of the angle of elevation. In cases where the radiated field varies with the bearing angle, the equivalent uniformly distributed field can be obtained by taking advantage of the fact that the radius of the circle

[1] Still another method is described by W. W. Hansen and J. G. Beckerley, Concerning New Methods of Calculating Radiation Resistance Either With or Without Ground, *Proc. I.R.E.*, vol. 24, p. 1594, December, 1936.

[2] Examples of such integrations are to be found in the following references: Ballantine, *loc. cit.;* Carter, Hansell, and Lindenblad, *loc. cit.*

enclosing the same area as does the actual distribution when plotted on polar paper represents the mean effective field strength. The *second step* is to plot the average effective field strength as a function of angle of elevation on polar paper and then multiply the resulting curve by $\sqrt{\cos \theta}$. The multiplication by $\sqrt{\cos \theta}$ takes into account the fact that on approaching the vertical the distance swept through in traveling around the sphere through a 360° bearing-angle rotation is proportional to the cosine of the angle of elevation, and the square root sign is necessary because the area under a curve plotted in polar form is proportional to the square of the radius vector. The *final step* consists in evaluating the area under this curve. This area is proportional to the total radiated power and can either be evaluated in terms of watts or can be compared with the corresponding area obtained from a comparison antenna in order to obtain the relative gain. The details of the graphical method

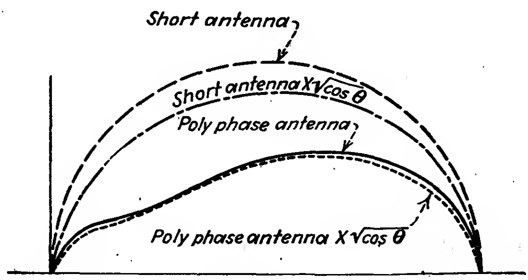


FIG. 407.—Curves used in obtaining the gain of the polyphase array of Fig. 401 compared with a short vertical antenna.

can be made clear by applying to a specific case such as the polyphase antenna system of Fig. 401. Here the average effective field strength along the horizontal is given by the radius of the circle shown in Fig. 401, which has the same area as the actual field distribution. When average effective values for different vertical angles are obtained in this way and plotted on polar paper, the result is as shown by the solid line in Fig. 407. Multiplication by $\sqrt{\cos \theta}$ gives the dotted curve, which incloses an area proportional to the total radiated power. For purposes of comparison the corresponding curves for a short vertical radiator (field strength proportional to $\cos \theta$) producing the same field strength along the horizontal are shown by the dash and dash-dot lines.

The merit of a directive system can be expressed in terms of the gain with respect to some comparison antenna. As explained in Sec. 130, this gain is the ratio of powers that must be supplied to the two antennas to produce the same field in the desired direction.[1] Thus in Fig. 407 the relative gain of the polyphase array as compared with the

[1] In making this comparison the antennas are assumed to have the same efficiencies.

short vertical radiator is equal to the ratio of areas under the dash-dot and dotted curves, and in this particular case is 1.70.

Evaluation of Power, Radiation Resistance, and Gain in Terms of Back Voltages Induced in the Antenna.—This method of analysis takes advantage of the fact that the fields produced by the antenna currents induce back voltages that are equivalent to an impedance against which the applied voltage can be considered as acting. With an isolated antenna, the resistance component of this antenna impedance is the radiation resistance, while the reactance component depends upon the tuning of the antenna and will be zero at resonance. When several antennas are in close proximity, the current in each causes a voltage to be induced in every other antenna, giving rise to mutual impedances such as exist in any other coupled circuit arrangement. A system of antennas can therefore be treated as a series of circuits each having a self-impedance, together with a mutual or transfer impedance to every other antenna present. The voltage and current relations existing in a system of radiators hence corresponds exactly with the voltage and current relations of a network,[1] and so can be written as

$$\left.\begin{aligned} E_1 &= I_1Z_{11} + I_2Z_{12} + \cdots + I_nZ_{1n} \\ E_2 &= I_1Z_{12} + I_2Z_{22} + \cdots + I_nZ_{2n} \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ E_n &= I_1Z_{1n} + I_2Z_{2n} + \cdots + I_nZ_{nn} \end{aligned}\right\} \qquad (244)$$

where

E_1, E_2, etc. = voltages applied to antennas 1, 2, etc.
I_1, I_2, etc. = currents flowing in antennas 1, 2, etc.
Z_{11}, Z_{22}, etc. = self impedances of antennas 1, 2, etc.
Z_{12}, Z_{2n}, etc. = mutual impedances between antennas denoted by subscripts.

The self impedances Z_{11}, Z_{22}, etc., and the mutual impedances Z_{12}, Z_{2n}, etc., can be calculated from the fields existing in the vicinity of the radiating wires and depend upon the geometry and type of current distribution involved. The actual details of the calculations are too involved to be presented here but can be found in the literature, together with formulas giving results for many practical cases.[2] Typical

[1] These equations are given in any book on network theory. For example see E. A. Guillemin, "Communication Networks," vol. 1, p. 156, John Wiley & Sons, Inc., New York.

[2] A comprehensive treatment of antennas from the point of view of self and mutual impedances is given by P. S. Carter, Circuit Relations and Radiating Systems and Application to Antenna Problems, *Proc. I.R.E.*, vol. 20, p. 1004, June, 1932. This paper gives a derivation of general formulas for the self and mutual impedances, and also formulas for a number of commonly used configurations such as the resonant V, harmonic-wire antennas, staggered and non-staggered parallel wires, colinear wires,

results of such calculations are shown in Fig 408. Here *a* gives the radiation resistance of an isolated wire as a function of length, while *b* gives the effect of spacing on the mutual impedance between two parallel half-wave radiators that are not staggered. It will be noted that as the spacing increases the mutual impedance diminishes rapidly and the real part may be either negative or positive according to the exact spacing.

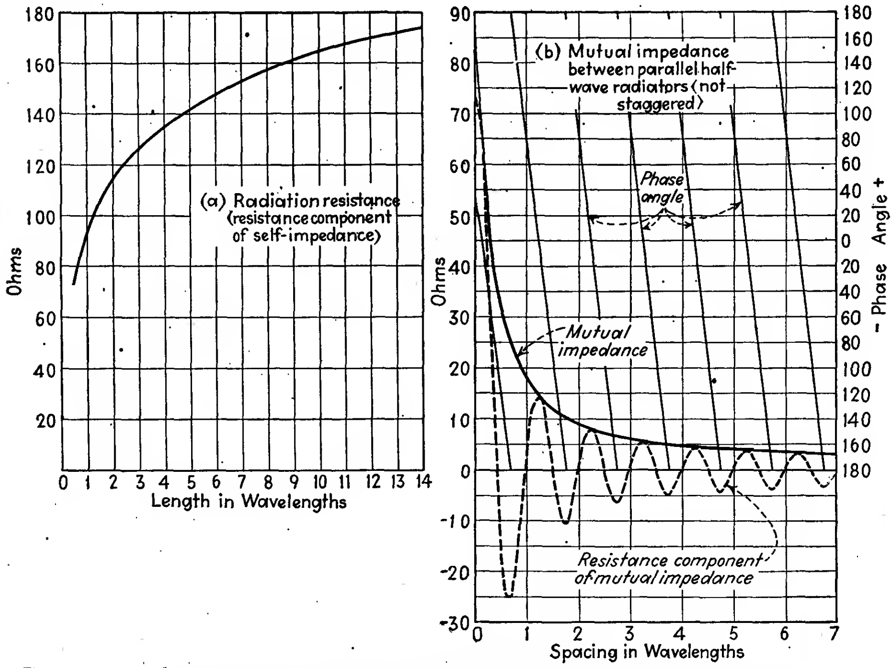


FIG. 408.—Radiation resistance of isolated wire antenna, together with mutual impedance between parallel half-wave radiators (not staggered).

The use of self and mutual impedances in the analysis of antenna problems can be made clear by typical examples:

Example 1.—Two half-wave radiators remote from earth are arranged broadside (same currents with same phase), and it is desired to determine the spacing for maximum gain. The solution is obtained by noting that the field produced in the desired direction is independent of spacing, so that the optimum spacing is that for which the

etc. Additional discussions of the same subject are given by A. A. Pistolkors, The Radiation of Beam Antennas, *Proc. I.R.E.*, vol. 17, p. 562, March, 1929; R. Bechmann, Calculation of Electric and Magnetic Field Strengths of Oscillating Straight Conductors, *Proc. I.R.E.*, vol. 19, p. 461, March, 1931; R. Bechmann, On the Calculation of Radiation Resistance of Antennas and Antenna Combinations, *Proc. I.R.E.*, vol. 19, p. 1471, August, 1931; F. H. Murray, Mutual Impedance of Two Skew Antenna Wires, *Proc. I.R.E.* vol. 21, p. 154, January, 1933.

power input is lowest. From Eq. (244) it is seen that the power supplied to the individual antenna is $I^2(R_{11} + R_{12})$, where R_{11} is the resistance component of the self impedance and R_{12} is the resistance component of the mutual impedance. From Fig. 408*a*, R_{11} for a half-wave antenna is seen to be 73.4 ohms, while R_{12} varies with spacing as shown in Fig. 408*b*. This causes the total antenna resistance, and hence the input power to the individual antenna, to vary as shown in Fig. 409. The minimum occurs at a spacing of 0.65 wave length, corresponding to an effective total resistance of 48.4 ohms for each antenna. The gain with this optimum spacing as compared with a single half-wave radiator is readily obtained by noting that, as compared with the single wire, the two antennas radiate twice as much field and hence four times the power in the desired direction, whereas the power input required is $2(48.4/73.4) = 1.32$ as great. The gain is accordingly $4/1.32 = 3.03$ times.

Example 2.—It is desired to determine the gain, in comparison with a single half-wave radiator, of a broadside array consisting of six half-wave radiators remote from

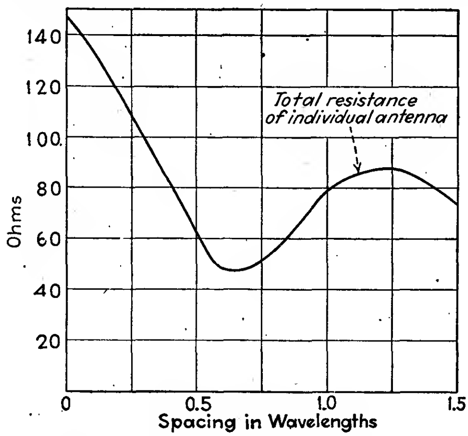


FIG. 409.—Variation of antenna resistance with spacing for two broadside half-wave radiators.

earth and spaced a half wave length apart. The six broadside antennas radiate 6 times the field and hence 36 times as much power in the desired direction as does the comparison antenna when both carry the same current. The power supplied to the comparison antenna is proportional to its radiation resistance of 73.4 ohms, while the power supplied to the broadside array is proportional to the sum of the radiation resistances of the individual wires plus the resistance components of all the possible combinations of mutual impedances. The total resistance of the end wire is therefore the radiation resistance plus the resistance component of the mutual impedances between the end wire and every other wire, and so from Fig. 408 is $73.4 - 15.5 + 3 - 0 + 0 - 0 = 60.9$ ohms. In similar fashion the total resistance of the next-to-end wires is $73.4 - (2 \times 15.5) + 3.0 - 0 + 0 = 45.4$ ohms, and for the middle wires is $73.4 - (2 \times 15.5) + (2 \times 3.0) - 0 = 48.4$. The power that must be supplied to the array must therefore be $2(60.9 + 45.4 + 48.4)/73.4 = 4.22$ times as great as that required by the comparison antenna, so that the relative gain of the array is $36/4.22 = 8.5$.

Another example showing how the behavior of parasitic antennas can be calculated from a knowledge of the self and mutual impedances is given in the next section.

Effect of Ground on the Energy Relations.—The methods of analyzing energy relations can be readily extended to include the effect of a perfect earth. When the analysis is based upon the radiated fields, the effect of the ground reflection upon the fields is taken into account by assuming the usual image antenna, but the integration for total radiated energy is carried out only over the hemisphere above the earth's surface, instead of over the complete spherical surface as with an isolated antenna. When the analysis is carried out in terms of mutual and self impedances, the mutual impedances between the image and the actual antenna elements are calculated, and the resulting effect upon the voltage, current, and power relations existing in the actual antenna then represent the effect of the ground.

In the case of an imperfect ground the situation is complicated since the radiated energy is less than the energy supplied to the antenna by the amount of the earth's losses. The principal effect of the earth losses is to absorb energy that would otherwise be radiated, without appreciably altering the impedance of the antenna.[1] Gain calculations are not affected appreciably by earth losses provided the actual and comparison antennas are in approximately the same relation to ground.

Formulas for Calculating Gain.—The formulas for calculating gain are too complicated and too specialized to warrant reproducing here but can be found by referring to the papers of Southworth, Sterba, Carter, Ballantine, and Carter-Hansell-Lindenblad, already referred to.

135. Directive Antenna Systems Employing Reflectors and Parasitic Antennas. *Parasitic Antennas.*—Directivity is sometimes obtained by the use of antennas that react upon the main antenna only through mutual impedance. The fundamental principles involved in such parasitic antennas can be understood by considering a simple example, such as a parasitic half-wave vertical antenna placed in the vicinity of a vertical half-wave radiator. The current I_1 in the radiating antenna induces a voltage in the parasitic antenna which according to Eq. (244) is

$$\text{Voltage induced in parasitic antenna} = -I_1 Z_{12}$$

The resulting induced current that flows is equal to this induced voltage divided by the self impedance Z_{22} of the parasitic antenna, so that

$$\text{Current in parasitic antenna} = I_2 = \frac{-I_1 Z_{12}}{Z_{22}}$$

Hence the magnitude and phase of the current in the parasitic antenna depend upon the spacing of the antennas (which determines the mutual

[1] This is substantiated by what literature is available dealing with the effect of earth losses on the power relations. Thus see W. L. Barrow, On the Impedance of a Vertical Half-wave Antenna above an Earth of Finite Conductivity. *Proc. I.R.E.*, vol. 23, p. 150, February, 1935.

impedance Z_{12}) and the tuning of the parasitic antenna (which determines the self impedance Z_{22}). The directive patterns that result for several typical conditions are shown in Fig. 410. It is seen that, when the spacing between the antennas is not too great, the parasitic antenna produces a marked directive effect because a large voltage is induced in it by the radiating antenna. The tuning of the parasitic antenna also has a pronounced effect since it alters both the magnitude and phase of the induced currents.

The most common use made of parasitic antennas is to produce a unidirectional effect in antenna arrays. Thus in arrays such as illustrated in Figs. 389 and 390 the rear curtain is very commonly excited parasitically. The optimum spacing for unidirectional action under these

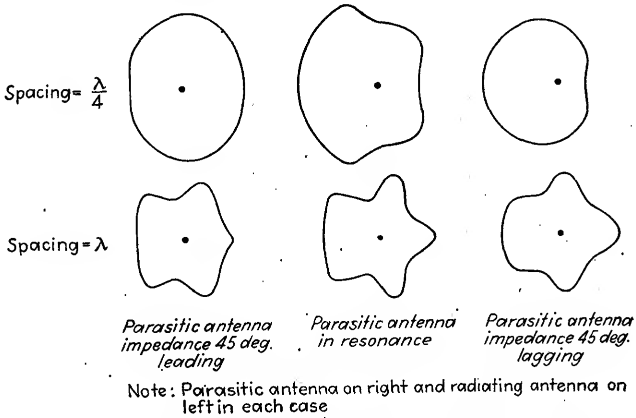


FIG. 410.—Directive patterns obtained with simple parasitic antenna under various conditions.

circumstances is approximately but not exactly a quarter of a wave length, and the parasitic antenna must be tuned so that its radiation is in the proper phase to produce the desired results.

Reflectors.—Another method of obtaining directivity is by the use of a metallic reflector, which in the case of a transmitting antenna concentrates the radiation in the desired direction in much the same manner as a mirror reflects light, and in the case of reception gathers in the energy from a wave coming from the desired direction and focuses it on the receiving antenna. Reflectors for this purpose are usually constructed of copper because this material has such high conductivity that it acts as a practically perfect reflector for radio waves.

Practical reflectors are shown in Fig. 411. In the parabolic types the antenna is placed at the focus and the concentrating action is similar to that obtained with a searchlight. The plane-surface type of reflector is used primarily to obtain a unidirectional characteristic with a broad-

side array. It replaces the reflecting antenna system otherwise required and to be effective it must be of ample dimensions.

The use of reflectors to obtain directivity is in practice limited to very short wave lengths, where the distance corresponding to a wave length is small enough to permit reflecting structures of practical size.

136. Radio-frequency Transmission Lines and Impedance Matching Systems.—The output of a radio transmitter is usually delivered to the

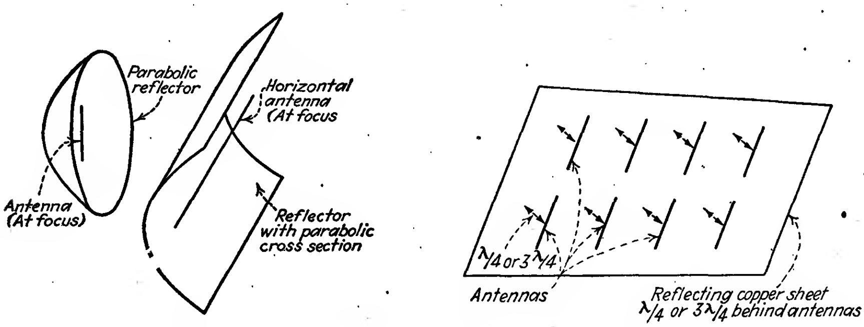


FIG. 411.—Examples of conducting reflectors for producing directive characteristics.

antenna by means of a transmission line. Transmission lines are also used to connect receivers to receiving antennas, particularly directive antennas, special all-wave antennas, and apartment-house antenna systems. By proper design the energy can be transmitted to distances of the order of ½ to 1 mile with only moderate loss.

The most common types of lines are the two-wire open-air line illustrated in Fig. 412a, the concentric-conductor line of Fig. 412b, the four-

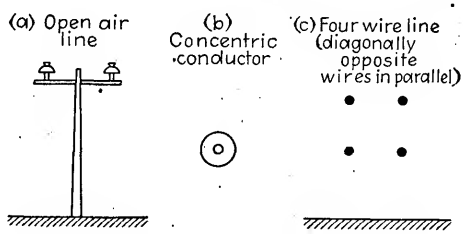


FIG. 412.—Common types of transmission lines.

wire balanced line of Fig. 412c, and a twisted pair composed of lamp cord or other insulated conductor. Lines used with transmitting antennas are usually of the two-wire open-air or concentric-conductor types. The open-wire line is commonly used because of its low cost, but the concentric-conductor type is less subject to weather, particularly when buried and filled with dry nitrogen gas under slight pressure to prevent collection of moisture from "breathing."

The requirements for lines used with receiving antennas depend upon the circumstances. Where it is merely a matter of transmitting the energy from antenna to receiver, any type of line is satisfactory, but with a directive receiving antenna it is very important that the transmission line abstract no energy whatsoever from passing waves. When there is abstraction of energy, one of the main objects of the directive receiving antenna will be lost because the antenna effect of the line will not be highly directive and so will cause the line to pick up noise and interfering signals to which the antenna itself is not responsive. The two-wire open-air line is not particularly satisfactory under such conditions because of the difficulty of maintaining perfect balance with respect to ground. The concentric transmission line, particularly when buried, is practically immune from stray pick-up but is expensive. A reasonably satisfactory and inexpensive compromise is furnished by the four-wire transmission line illustrated in Fig. 412*c*, which consists of four wires arranged at the corners of a square approximately 1 to 2 in. to a side and with the diagonally opposed wires connected in parallel. This arrangement is simple and when regularly transposed is very effectively balanced against extraneous disturbances.

Twisted pair lines are often used in low-power transmitters and in lines associated with receivers. Such lines are not susceptible to stray pick-up because of the small spacing and frequent twisting, but are suitable only for short lengths because of their high losses.[1]

Resonant and Non-resonant Lines.—Transmission lines may be operated in such a manner as to be resonant or non-resonant, according to the load impedance at the receiving end of the line. If the receiver load is a resistance equal to the characteristic impedance (*i.e.*, a resistance equal to $\sqrt{L/C}$), then voltage and current die away uniformly as one recedes from the sending end of the line, and have a distribution that follows an exponential law as shown at Fig. 33*c*. There are then no resonances irrespective of the line length; the voltage and current at every point along the line are in phase with each other; the ratio of voltage to current at any point is the characteristic impedance $Z_0 = \sqrt{L/C}$; and the phase drops back uniformly at a rate of 360° per wave length as the receiver is approached.[2]

When the load impedance does not equal the characteristic impedance of the transmission line, resonance effects are produced, as illustrated in

[1] See C. C. Harris, Losses in Twisted Pair Transmission Lines at Radio Frequencies, *Proc. I.R.E.*, vol. 24, p. 425, March, 1936.

[2] In the case of lines with air insulation, the distance corresponding to a wave length on the line is almost exactly equal to the wave length of the corresponding radio wave. However, in the case of lines having appreciable dielectric insulation, as, for example, a twisted pair, the distance representing a wave length on the line is correspondingly less.

Fig. 33, and the resonances will be very pronounced if the load impedance differs appreciably from the characteristic impedance. With resonant lines the voltage and current follow cyclical variations which repeat every half wave length, and there is a large phase difference between voltage and current except near points where the voltage or current passes through a minimum. In traveling from generator toward load there is a progressive phase lag of 360° per wave length, but, instead of taking place at a uniform rate as in the case of non-resonant lines, most of the shift takes place in the vicinity of the voltage and current minima.

Non-resonant lines transmit energy at unity power factor and so are more efficient than resonant lines, and also subject the line insulation to less voltage in proportion to power transmitted. Hence non-resonant lines are always used in association with high-power transmitters and in all cases where the transmission distance is appreciable. The resonant lines are suitable only with low-power transmitters where the transmission distance is quite short, as in the case of amateur, airplane, and similar equipment.

Properties of Transmission Lines.[1]—The most important characteristics of a transmission line are the characteristic impedance Z_0 and the attenuation factor α. The characteristic impedance is determined by the line construction and at radio frequencies can for all practical purposes be considered to be a resistance defined by the equation[2]

$$Z_0 = \sqrt{\frac{L}{C}} \text{ ohms} \tag{245a}$$

where L and C are the inductance and capacity, respectively, per unit length of line. For the two-wire and concentric-conductor transmission lines, substitution of the usual formulas for inductance and capacity gives

$$\left.\begin{array}{l}\text{Characteristic impedance}\\ \text{of two-wire line}\end{array}\right\} = 276 \log_{10} \frac{b}{a} \text{ ohms} \tag{246a}$$

$$\left.\begin{array}{l}\text{Characteristic impedance}\\ \text{of concentric cable}\end{array}\right\} = 138 \log_{10} \frac{b}{a} \text{ ohms} \tag{246b}$$

[1] For further information see E. J. Sterba and C. B. Feldman, Transmission Lines for Short Wave Radio Systems, *Proc. I.R.E.*, vol. 20, p. 1163, July, 1932. This paper gives an excellent discussion of the electrical properties of two-wire and concentric-conductor lines, constructional problems, correlation of measured and predicted behavior, etc.

[2] The exact formula for the characteristic impedance is

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{245b}$$

where R and G are the resistance and conductance per unit length of line. At high frequencies ω is so large that R and G can be neglected, thereby resulting in Eq. (245a).

In these equations b represents the spacing between conductors of the two-wire line and the inner radius of the outer conductor of the concentric cable, while a represents the radius of the conductor in the two-wire line and the outer radius of the inner conductor in the case of the concentric line. With the usual construction the characteristic impedance is commonly of the order of 600 ohms with a two-conductor open-air line, and in the neighborhood of 75 ohms for the usual concentric conductor with air dielectric. With a twisted-pair transmission line the characteristic impedance is commonly of the order of 125 ohms, but will depend greatly upon the construction.

The attenuation constant of a transmission line depends upon the resistance, leakage, and characteristic impedance, and at radio frequencies can be expressed by the equation

$$\text{Attenuation constant} = \alpha = \frac{R}{2Z_0} + \frac{GZ_0}{2} \tag{247}$$

where R and G are the resistance and leakage per unit length of transmission line and Z_0 is the characteristic impedance.

In the usual two-wire and concentric-conductor lines most of the insulation is air, so that the leakage is substantially zero except when the insulators are wet, and the conductor resistance accounts for nearly all the attenuation. This resistance is determined largely by skin effect, and can be calculated by the following formulas:[1]

$$\left.\begin{matrix}\text{Resistance of}\\ \text{concentric conductor}\end{matrix}\right\} = 41.6\sqrt{f}\left(\frac{1}{a} + \frac{1}{b}\right) \times 10^{-9} \text{ ohms per centimeter} \tag{248a}$$

$$\left.\begin{matrix}\text{Resistance of two-}\\ \text{wire line}\end{matrix}\right\} = 83.2\frac{\sqrt{f}}{a} \times 10^{-9} \text{ ohms per centimeter} \tag{248b}$$

The frequency is in cycles, the dimensions are in centimeters, and a and b have the same meaning as in Eq. (246). It will be noted that the resistance is directly proportional to the square root of the frequency and inversely proportional to the physical size of the transmission line. In a concentric conductor having a fixed radius b of the outer conductor, the lowest resistance will be obtained when the inner conductor is such that $b/a = 3.6$.

The behavior of a non-resonant line can be summarized by the following equations:

[1] The two-wire formula neglects proximity effect, and so is about 15 per cent low for $b/a = 4$, and less than 2 per cent low for $b/a > 12$.

$$\left.\begin{aligned} E &= E_s\epsilon^{-\alpha l} \\ I &= I_s\epsilon^{-\alpha l} \\ \frac{E}{I} &= Z_0 = \text{characteristic impedance} \\ &\left.\begin{matrix}\text{Phase shift of } E \text{ and} \\ I \text{ per unit length}\end{matrix}\right\} = \beta = \omega\sqrt{LC} = \frac{2\pi}{\lambda} \\ &\text{Power loss} = 8.686\alpha l \text{ db} \end{aligned}\right\} \tag{249}$$

where

$\omega = 2\pi f$

λ = length along line corresponding to one wave length

E_s and I_s = voltage and current at the generator end of the line

E and I = voltage and current at a distance l from the generator.

The losses in a properly constructed line can be made very small. Thus by combining Eqs. (247), (248*b*), and (249) it is found that a two-wire line made of No. 4 copper wire and having a characteristic impedance of 600 ohms introduces an attenuation of 0.24 db per thousand feet at a frequency of 10 mc.

Radiation from Transmission Lines.—A two-wire transmission line radiates very little energy because the close proximity of the two conductors carrying current in opposite directions very nearly cancels the radiated field. Analysis shows that in the case of a perfectly balanced two-wire line the total radiation, including that of the terminal connections, is twice the radiation that would be obtained from an elementary antenna having a length equal to the transmission line spacing and carrying a current equal to the line current. With resonance on the line the radiated power is greater because of the larger line currents.[1] When a two- or four-wire line is not balanced with respect to ground, there is a current component that travels out over the wires acting in parallel and returns by way of the ground. The radiation under such conditions will depend upon the amount of unbalance and the height of the transmission line above earth, and will in general tend to be large because the height is much greater than the spacing between wires.

A concentric transmission line radiates no energy under ordinary conditions because the outer conductor acts as a substantially perfect shield.

[1] The power radiated from a non-resonant two-wire line.

$$\text{Power radiated} = 160I^2\left(\frac{\pi a}{\lambda}\right)^2 \text{ watts} \tag{250}$$

where I is the line current (assumed constant) and a/λ is the spacing in wave lengths. For a resonant line the radiated power depends on the amount of resonance, but is greater than given by Eq. (250) by a factor that is very roughly $(I_0/I)^2$, where I_0 is the maximum current in the resonant line. For further information see Sterba and Feldman, *loc. cit.*

The tendency of a transmission line to radiate when carrying currents is also a measure of the extent to which the transmission line will act as an antenna in the reception of radio waves. Accordingly, a receiving antenna transmission line that is to pick up the minimum possible energy is designed in the same way as a line that is to radiate the minimum possible energy. With open-air lines this means non-resonant operation, careful balance with respect to ground, and close spacing. A concentric conductor line is particularly suitable for receiving purposes because the

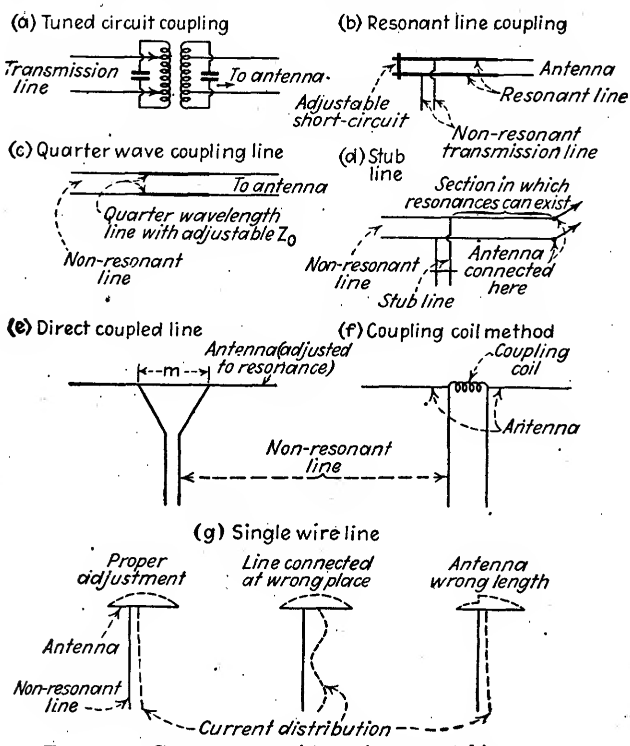


FIG. 413.—Commonly used impedance-matching systems.

outer conductor acts as a practically perfect shield against radio-frequency waves.

Impedance-matching Systems for Non-resonant Lines.—In order to avoid resonances in transmission lines associated with transmitters, it is necessary that the antenna be coupled to the line in such a way that the effective load impedance which is offered to the line is a resistance equal to the characteristic impedance of the line. Typical methods for doing this are illustrated in Fig. 413. At Fig. 413*a* a tuned circuit is used to provide the impedance matching. A resistance load is obtained by proper tuning, while the magnitude of the load can be adjusted to the

correct value either by varying the mutual inductance or by varying the fraction of the coil across which the line is connected.

In the resonant coupling line at Fig. 413*b* a resistance impedance is obtained by adjusting the short circuit on the coupling line to give resonance in conjunction with the antenna. The resulting load offered the non-resonant line is then resistive and has a magnitude determined by the point of connection. The length of resonant line should be approximately a quarter wave length when the antenna resistance is higher than the non-resonant line impedance, and a half wave length when the antenna resistance is lower. Arrangements employing resonant coupling lines are shown in Figs. 397 and 399*a*.

The arrangement of Fig. 413*c* makes use of the fact that in a quarter-wave-length coupling line $Z_s = Z_0^2/Z_a$, where Z_s is the impedance that the non-resonant transmission line sees when looking toward the matching line, Z_0 is the characteristic impedance of the matching line (a resistance), and Z_a is the antenna impedance. By adjusting the antenna so that it is in resonance, Z_a becomes a resistance, and the impedance Z_s, which serves as the load for the non-resonant line, is likewise a resistance which can be made the desired value by varying the spacing between wires of the matching line to vary Z_0. At low frequencies where a quarter wave length is inconveniently long, it is possible to use an artificial line designed as explained in the footnote on page 541.[1]

In the stub-line arrangement of Fig. 413*d* the transmission line is connected directly to the antenna and a suitable shunt reactance, commonly in the form of a short length of transmission line (*i.e.*, a stub line), is used to eliminate resonances between the shunting point and the generator. This shunting reactance must be placed at a point such that, when the impedance looking toward the antenna is paralleled by the shunt reactance, the impedance of the combination is a resistance equal in magnitude to the characteristic impedance of the line. The exact details can be worked out theoretically and lead to the results that are presented graphically in Fig. 414.

Several other coupling methods sometimes employed are shown at *e*, *f*, and *g* in Fig. 413. At *e* the transmission line is directly coupled to the antenna, the matching being obtained by making the antenna length the exact value required for resonance and then connecting the two wires of the line symmetrically with a spacing *m* such as to give the

[1] Information on the design of such lines for broadcast transmitters is also given by Ralph P. Glover, R-f Impedance Matching Networks, *Electronics*, vol. 9, p. 29, January, 1936; Carl G. Dietsch, Terminating Concentric Lines, *Electronics*, vol. 9, p. 16, December, 1936. An excellent discussion of impedance-matching networks in general is given by W. L. Everitt, Output Networks for Radio-frequency Power Amplifiers, *Proc. I.R.E.*, vol. 19, p. 725, May, 1931.

required impedance match. At f the antenna is made shorter than is required for resonance and is then tuned by means of a small coil to which the transmission line is directly coupled. The single-wire transmission line shown at g is a modification of Fig. 413e in which the ground supplies a return circuit that is completed through the capacity of the antenna

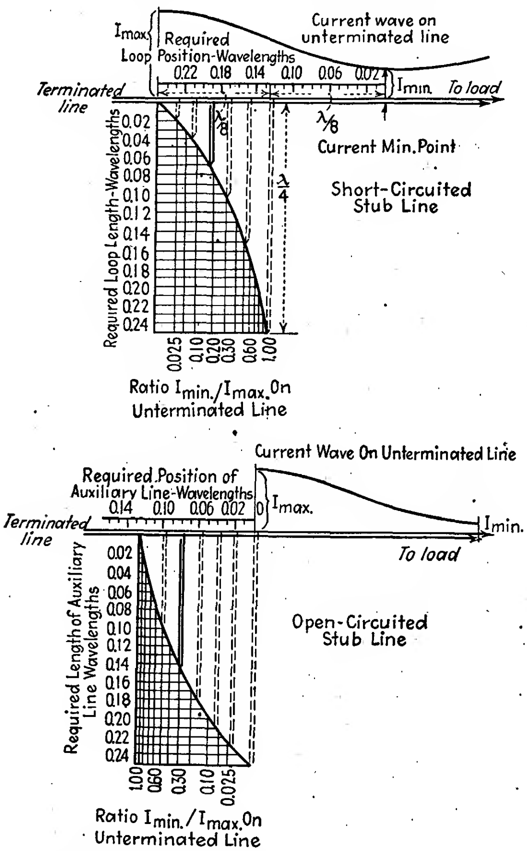


FIG. 414.—Curves giving design information on stub lines. These curves give the required position and length of stub line in terms of the ratio of minimum to maximum current on the line before the stub line is added. The location is indicated with respect to a current maximum.

to ground. The proper adjustment of a single-wire transmission line can be determined by observing the relative current distribution in the line and antenna as indicated.[1] The single-wire line radiates much more energy than the corresponding two-wire line, but is often used because of its simplicity.

[1] See W. L. Everitt and J. F. Byrne, Single-wire Transmission Lines for Short-wave Antennas, *Proc. I.R.E.*, vol. 17, p. 1840, October, 1929.

When it is desired to avoid resonances in a transmission line associated with a receiving antenna, it is necessary that the input impedance of the receiver be a resistance equal to the characteristic impedance of the line, since in this case the antenna is the source of energy and the receiver acts as the load. Failure to match the antenna to the transmission line prevents the receiving antenna from delivering the maximum possible amount of energy to the transmission line, but does not cause resonances.

Phasing Systems.—In antenna arrays it is necessary to adjust carefully the relative phases of the various antennas. There are two basic

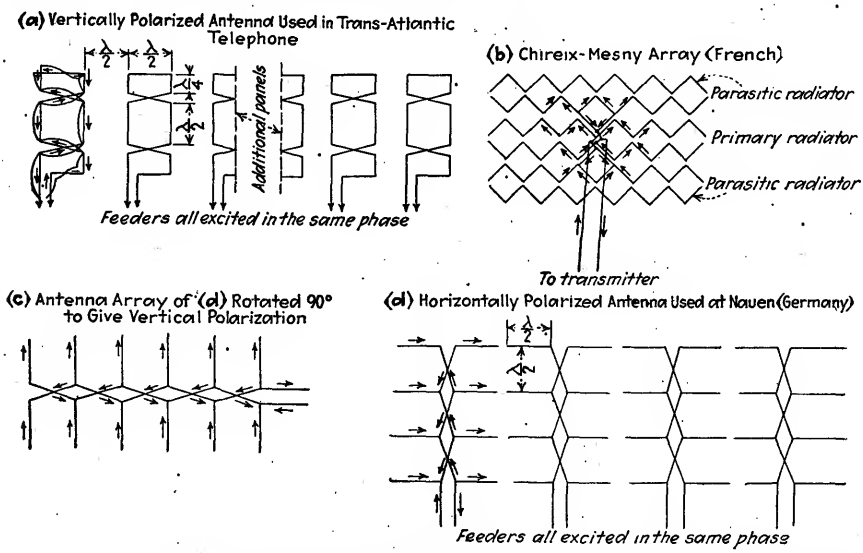


FIG. 415.—Commercial designs of broadside antenna arrays. The reflector antenna that is always used to give a unidirectional characteristic is omitted in the figures for the sake of clearness.

methods of accomplishing this. The first makes use of the fact that a non-resonant transmission line in open air has a uniform phase shift of 360° per wave length. Hence any desired phase difference can be obtained by the use of a non-resonant transmission line of suitable length. Thus in the antenna system of Fig. 397 the 2¼-wave-length non-resonant line connecting antenna and reflector makes the relative phase difference 2¼ cycles, *i.e.*, 90°. Phasing of end-fire antenna arrays is always accomplished by the aid of a non-resonant line.

The second method of phasing makes use of the fact that in a resonant line the phase shifts 180° every half wave length (even though the shift does not take place uniformly). This method is commonly used in phasing broadside antenna arrays. Typical detailed illustrations are shown in Fig. 415. At *c* and *d* the radiators are connected to the reso-

nant line at voltage loops and are spaced a half wave length apart so that by connecting successive radiators to alternate sides of the line all the radiators are excited in the same phase. At *a* the antenna is formed by folding the wire in such a manner that the vertical sections all carry current in the same phase while the horizontal sections serve as half-wave non-radiating resonant lines that produce a phase reversal. At *b* the antenna wires serve simultaneously as radiators and transmission lines by being so bent that the currents along any diagonal are in the same phase. The diagonals in opposite directions then combine to give the effect of a vertical polarized array.

Experimental Determination of Current Distribution along Transmission Lines and Antennas.—The way in which a transmission line or

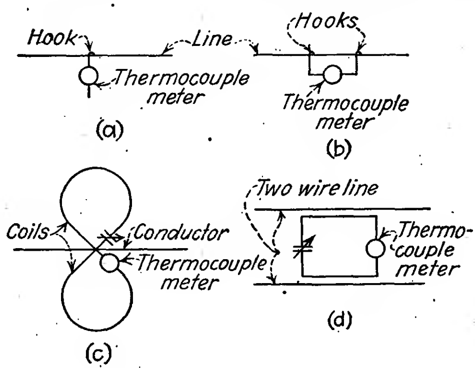


Fig. 416.—Devices for measuring the relative current distribution along a transmission line or antenna.

antenna is operating can ordinarily be deduced by observing the current distribution. In the case of a transmitting antenna the current distribution can be measured by coupling a sensitive thermocouple instrument to the line, using one of the arrangements in Fig. 416. In the case of receiving antennas the thermocouple can be replaced by a portable receiver and the antenna can be excited by means of a portable oscillator adjusted to produce radio waves of the appropriate frequency.

Radio-frequency Distributing Systems.[1]—It is sometimes desired to operate a number of receivers from the same antenna, employing a transmission line to carry the energy to the various locations involved. Such an arrangement is particularly desirable in apartment houses in large cities, since it permits the installation of one efficient antenna on the roof remote from man-made noises and where the signals are strong.

The most satisfactory arrangement of this type employs a non-resonant transmission line to which the various receivers are coupled

[1] Further information on this subject is given by F. X. Rettenmeyer, Radio-frequency Distributing Systems, *Proc. I.R.E.*, vol. 23, p. 1286, November, 1935.

through an attenuation unit that introduces about 30 db loss. The loss in the attenuator is compensated for either by employing an individual buffer amplifier between each receiver and the line or by means of a wide-band amplifier connected between the antenna and the transmission line. With an arrangement of this sort satisfactory reception can be obtained with receivers having only ordinary sensitivity, and interaction between receivers is practically eliminated because effects originating in a particular receiver, such as local oscillations, are attenuated 60 db before reaching the input of a second receiver.

In simple installations where only a few receivers are involved, the attenuating network and amplifier are sometimes omitted. In such cases the receiver is provided with a high impedance input circuit which is connected directly across the transmission line.

137. Practical Transmitting Antennas.—The principal considerations involved in the design of transmitting antennas are the directivity, efficiency, and cost. These matters become of increasing importance as the power of the transmitter is increased since the expenditures for improving the antenna system are then easier to justify.

Short-wave Antennas.—Short-wave directive antennas find their chief use in long-distance high-power point-to-point communication. For this purpose it is desirable to concentrate the energy in a well-defined beam which is directed toward the receiving point at a vertical angle that is of the order of 10 to 25 deg. above the horizontal. The rhombic and resonant V antennas are usually preferred because they give good directivity, are easily tuned, and can be supported upon inexpensive wooden poles. The rhombic antenna has the additional advantage that it will operate satisfactorily over a frequency range approximately 2½ to 1 with no readjustment, and so can be used for both the day and night frequencies. Before the resonant V and rhombic antennas were developed, broadside arrays such as illustrated in Figs. 389 and 415 were generally employed, but these have lost favor because they are difficult to tune, require expensive supporting structures, and give only slightly better directivity.

Practical experience indicates that for best results the angular spread of the beam in the horizontal plane should be of the order of 10 deg. while in the vertical plane a spread of perhaps 15 deg. centered at a vertical angle of 15 to 20 deg. is optimum. With less sharply defined beams energy is wasted by being sent in directions where it has no chance to reach the receiver, while, if greater directivity is used, the beam will sometimes be deflected sufficiently from the great-circle route to make the received signals weaker than when a broader beam is used. Interpreted in terms of practical antenna arrays, the amount of gain that can be usefully employed is of the order of 100.

Non-directional radiators consisting of a vertical or horizontal half-wave antenna are commonly used for amateur, marine, and short-wave broadcast work, where it is desired to transmit signals in all directions and for varying distances. Typical examples of such antenna systems, together with typical transmission line arrangements, are shown in Fig. 416*a*.

The polarization of the transmitting antenna is not important since the ionosphere causes the received wave to contain both vertical and horizontal components irrespective of the polarization at the transmitter. Horizontal polarization is usually preferred, however, because the supporting structures required are generally not so high as for vertical radiators with the same vertical directivity, and furthermore with vertical radiators it is probable that the Sommerfield ground wave in the vicinity of the transmitter absorbs an appreciable fraction of the radiated energy.

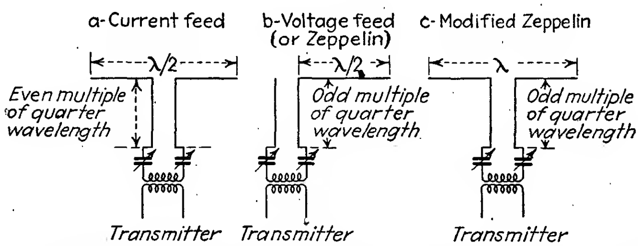


FIG. 416*a*.—Representative types of resonant transmission lines for delivering energy to antennas.

The proper procedure for tuning an antenna depends upon the arrangement. In the case of arrays involving many radiators, such as illustrated in Fig. 389, the length of the individual parts are adjusted until the current distribution as experimentally observed has the desired character. In other cases where the antenna is associated with a resonant line such as Figs. 397, 399, and 416*a*, it is possible to tune the antenna and the resonant line as a unit. At Figs. 397 and 399 this is done by adjusting the position of the short-circuiting bar, while in arrangements of the type shown in Fig. 416*a* the tuning can be accomplished by means of the variable condensers in series with the line, provided the antenna and transmission-line lengths are approximately correct to begin with. In cases such as Fig. 413*e* where the antenna must offer a resistance impedance, it is necessary to adjust the antenna length very accurately to the proper value.

In general it will be found that the antenna length required for resonance is usually less than the corresponding length in wave lengths measured in free space. . This is because of end effects and mutual coupling in antenna elements. Thus in the case of a half-wave antenna it is

necessary to make the wire about 5 per cent shorter than a true half wave length, while in an array composed of a multiplicity of half wave length antennas the elements sometimes must be as much as 10 per cent shorter.

When a reflector system is employed to give a unidirectional characteristic, the tuning procedure in the case of a parasitic antenna is first to adjust the radiating antenna to give the desired current distribution, after which the element lengths in the reflecting antenna are adjusted until the backward radiation as observed upon a suitably located radio receiver is a minimum.

In cases where both antenna and reflectors are excited, the antenna and reflector lengths are adjusted to give the desired current distribution, and the relative phasing is maintained by coupling the two together with a non-resonant line of the proper length. Under such conditions it is desirable to have the reflector spaced at least ¾ wave lengths from the antenna in order to prevent excessive interaction during the adjustment process.

The efficiency of short-wave antennas is very high because of the high radiation resistance of the antenna in proportion to size and because the antenna is ungrounded. Even including the effect of ground losses, the over-all efficiency is commonly well in excess of 80 per cent, and most of the losses that do occur are the result of ground imperfections.

Ultra-high-frequency Antennas.—Antennas for use with ultra-high-frequency transmitters are of the same general character as those used at short-waves. However, since the length corresponding to a wave length is less, it is possible to obtain greater directivity with antennas of reasonable proportions.

In ultra-high-frequency broadcasting it is desirable to have an antenna located at a high elevation and possessing considerable vertical but no horizontal directivity. The necessary elevation is usually obtained by locating the antenna upon a flagpole of a high building, while vertical directivity can be obtained by stacking colinear half-wave radiators or by a turnstile array of the type illustrated in Fig. 400. In point-to-point communication with ultra-high frequencies it is possible to make use of almost unlimited directivity, but the beam must be directed toward the receiver (not at an angle above the horizontal as in the case of short waves) and must have sufficient elevation above earth to minimize destructive interference from the reflected ground wave.

Transmitting Antennas for Broadcast Frequencies.[1]—Nearly all the more recent broadcast antennas are self-supporting steel towers having a

[1] For further information on this subject see H. E. Gihring and G. H. Brown, General Consideration of Tower Antennas for Broadcast Use, *Proc. I.R.E.*, vol. 23, p. 311, April, 1935; G. H. Brown, A Critical Study of the Characteristics of Broadcast

height either approximately a quarter wave length or slightly in excess of a half wave length, and serving as vertical radiators. Such an antenna can be excited by insulating the base of the tower from ground and applying the exciting voltage between the base and ground, or by making use of a loop formed by the ground and a wire attached to the tower above ground at a height corresponding to approximately 20 per cent of the antenna height, as illustrated in Fig. 417*d*. This loop is made resonant to the transmitted frequency by means of the capacity *C*, and therefore carries a large circulating current which develops sufficient voltage across the section *h* to excite the remainder of the tower. The tower radiator has the advantage of simplicity, low cost, and good efficiency.

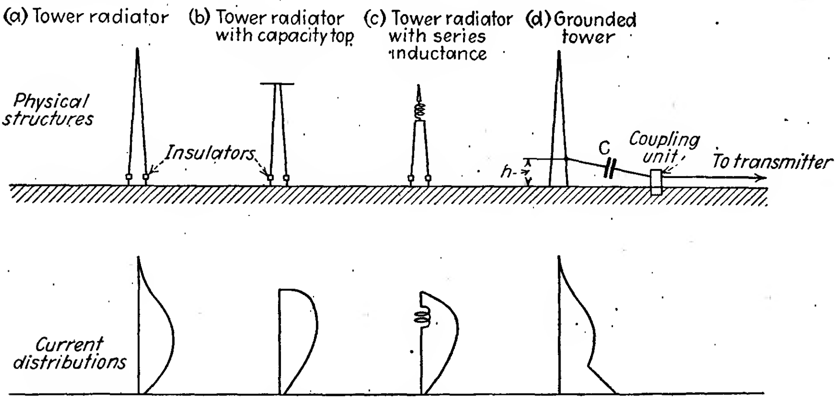


FIG. 417.—Different types of tower radiators, with current distribution for each.

The vertical directivity of a vertical radiator depends upon the height, as shown in Fig. 378. The optimum results for broadcast purposes are obtained when the height is of the order of 0.50 to 0.56 wave length, as this gives maximum concentration of energy along the ground consistent with suppression of the sky wave (see Sec. 122 for discussion of the desired directional characteristics). With heights much less than 0.5 wave length the directivity is largely lost, and heights greater than 0.6 wave length produce a strong high-angle sky wave.

The relatively great cost of vertical radiators with optimum height has led to the investigation of possible alternative arrangements. Among these are the polyphase array of Fig. 401, the capacity-top arrangement of Fig. 417*b*, and the sectionalized tower with series inductance shown at 417*c*. The polyphase array is capable of giving greater directivity than obtainable with a simple vertical radiator, but is also much more

Antennas as Affected by Antenna Current Distribution, *Proc. I.R.E.*, vol. 24, p. 48, January, 1936; A. B. Chamberlain and W. B. Lodge, The Broadcast Antenna, *Proc. I.R.E.*, vol. 24, p. 11, January, 1936.

expensive. The capacity top reduces slightly the height required for optimum operation, but the saving with practical structures is hardly worth bothering with. The sectionalized towers with series inductance make it possible to reduce the total height from 20 to 30 per cent without sacrifice in directive characteristics, but this is in part counterbalanced by the extra cost of sectionalizing the tower and by the energy losses in the inductance.[1] However, sectionalizing has the advantage of making it possible to adjust the effective height experimentally without involving expensive structural changes and so is sometimes used for this reason.

The amount of directivity obtainable with a vertical radiator about 0.54 wave length high approximates the optimum directivity for a 50-kw station operating in the frequency range of 750 to 1000 kc with average soil and receiving conditions. With high-conductivity soil, lower frequency,

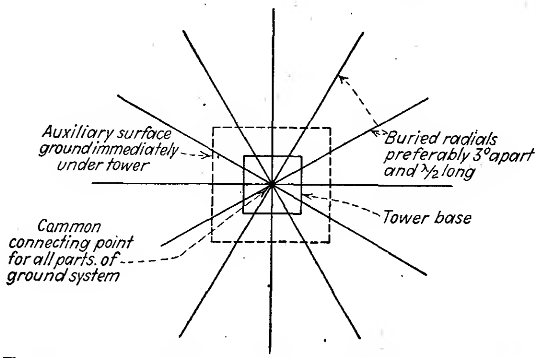


FIG. 418.—Typical ground system for vertical radiator.

favorable receiving conditions, or greater transmitter power, more directivity would ordinarily be useful if obtainable at reasonable cost. Under typical conditions a vertical radiator of optimum height roughly doubles the primary service area of a 50-kw station as compared with a short vertical radiator. A portion of this gain is the result of a stronger ground wave, while the remainder is due to the reduction of the high-angle sky wave.

Where a vertical radiator of optimum height is not practicable or cannot be justified economically, as in the case of transmitters with low power or low frequency, or when established airplane routes pass in the vicinity of the antenna, it is customary to use radiating towers having a height of approximately one-quarter wave length. Such an arrangement radiates a field that is almost exactly proportional to the cosine of the

[1] An analysis of the sectionalized tower is given by G. H. Brown, A Note on the Placement of the Coil in a Sectionalized Antenna, *Broadcast News*, no. 22, p. 14, October, 1936.

angle of elevation, and so has no more directivity than a short antenna with flat top, but is less expensive and more efficient.

The efficiency of a broadcast antenna depends largely upon the ground losses occasioned by currents flowing in the earth near the antenna. These losses can be made low by providing a ground system of buried wires as shown in Fig. 418, consisting of numerous radials having a length that is preferably not less than a half wave length. These radials provide a low-resistance path for the currents that flow out through the ground to charge the capacity between tower and ground, and they thereby reduce the effective ground resistance. With tower antennas it is customary to use between 90 and 120 such radials. These should, if possible, be about a half wave length long, since when appreciably shorter the ground losses increase greatly even when the antenna height is small.[1] With such a ground system the efficiency of a tower radiator is high.

A self-supporting tower insulated from ground has a relatively high localized capacity from the lower end of the tower to the earth immediately below. Losses from this source can be reduced to a negligible value by placing the tower insulators on posts from 6 to 10 ft. high in order to reduce the capacity to ground and by providing an additional grounding system in the immediate vicinity of the tower base. This auxiliary ground can be a screen laid on the earth's surface and extending out a short distance beyond the tower footings, or can be a system of closely spaced overhead radial wires bonded together and connected to the buried grounding system at one point.[2]

Calculation of the exact behavior of tower antennas is complicated by the fact that a self-supporting tower is tapered, and this causes the current distribution to deviate somewhat from sinusoidal. Experiments indicate that with the self-supporting tower radiator the current distribution is of the form shown in Fig. 417*a*. As a result of the reduced current in the top portion of a tapered tower, it is necessary that the tower have a slightly greater height when tapered than when of uniform cross section. Otherwise there is no essential difference in spite of the fact that a uniform-cross-section structure is sometimes believed to be capable of giving a performance that cannot be duplicated by the tapered

[1] Thus G. H. Brown, The Phase and Magnitude of Earth Currents near Radio Transmitting Antennas, *Proc. I.R.E.*, vol. 23, p. 168, February, 1935, shows that with a half-wave tower the maximum earth losses in a circular shell about the tower are at a distance approximately $\lambda/3$ distant from the tower base.

[2] The first tower radiators were specially designed to have a very small capacity from the lower end to ground, since it was assumed that this capacity was highly detrimental to the behavior. However, analysis shows that the ground capacity is harmful only if the ground losses in the immediate vicinity of the tower base are high, and the self-supporting type of structure is now generally accepted as being entirely suitable.

tower. The only advantage of the uniform cross section is that it enables one to predict the optimum height accurately by calculation (optimum is 0.53 wave length), while with a tapered tower the optimum height depends upon the rate of taper and there is not enough information available on these towers to permit an exact predetermination of the best height.

Simple antenna arrays are used to some extent to give horizontal directivity in broadcast work. Thus it is sometimes desired to concentrate most of the radiated energy toward the regions of maximum population or to avoid interfering with another station. These objectives can often be achieved by means of a simple array consisting of two or three radiators properly phased. Thus a broadside array consisting of three vertical radiators spaced one-half wave length apart will give a field strength approximately three times as great in the broadside direction as in the right-angle direction.

Transmitting Antennas for Long Waves.—At frequencies below the broadcast band the distance corresponding to a wave length is so great that it is impracticable to obtain directivity with efficiency and reasonable cost. The usual antenna system for the lower frequencies consists of a T, or inverted L, flat-top arrangement, as illustrated in Fig. 371*e* and *f*, supported by two towers. The flat top radiates very little energy but serves two very important functions. In the first place, if its capacity to ground is considerably greater than the capacity of the vertical wire, the current in the vertical wire will be substantially constant over its entire length. This increases the effective height of the antenna by giving more radiated energy in proportion to the antenna current and height, with the result that the antenna efficiency is improved. In the second place, the flat top reduces the antenna voltage, since with a large-capacity top it is possible to have a large current in the vertical section with a relatively low antenna voltage. This is important when the power is high and the frequency is low. Thus, when the power is of the order of 100 kw at a frequency of 20 kc, a flat-top structure 1 mile long and 500 ft. wide is required to keep the antenna voltages low enough to permit adequate insulation and avoid corona losses.

The efficiency of low-frequency antennas tends to be very small because of the low radiation resistance obtained with practical heights and because the power losses in the ground and in the antenna tuning coil tend to be large as a result of the large physical dimensions involved. The ground system is particularly important because, with a large flat top, currents must flow through the ground for a considerable distance. The preferred arrangement is to carry these ground currents by means of buried or overhead wires to the immediate vicinity of the point where these currents leave the ground to charge the capacity between antenna

and ground.[1] In this way the distance that must be traveled through the earth is greatly reduced.

Miscellaneous.—Steel towers, guy wires, and other conductors in the vicinity of an antenna tend to alter the directive characteristics and so must be located with considerable care. Steel supporting towers should not have a length that makes them resonant to the frequency being transmitted, and they should be located where the coupling to the antenna system is a minimum. Guy wires should not be too numerous, and must be broken up into lengths that are a small fraction of a half wave length by means of insulators.

Transmission lines used with transmitting antennas should approach the antenna at right angles so that there will be a minimum of coupling between the antenna and line.

In some locations trouble is encountered from sleet. This can be avoided by arranging the antenna system so that a 60-cycle heating current can be passed through the antenna wires to prevent the formation of sleet. It is desirable that the arrangements be such that the heating current can be applied while the antenna is carrying on its normal radiating functions.

Antennas for use with airplane transmitters introduce special problems. In some cases the antenna is strung from wing tip to tail, or from a short vertical pole to wing tip, etc., but such antennas have a low effective height and tend to be rather inefficient even with short waves. A more satisfactory arrangement from the point of view of efficiency is a trailing wire reeled out through a hole in the rear end of the fuselage. This wire trails behind the plane with little or no weight at its end, and is made weak enough so that, if the pilot forgets to reel in the wire before landing, it will break without producing any serious reaction on the plane.

138. Receiving Antennas.—The main considerations involved in receiving antennas are the amount of energy that the antenna can deliver to the receiver, the directivity, the cost, and the freedom from extraneous disturbances.

The energy that an antenna is able to deliver to a radio receiver depends upon the physical size of the antenna, the frequency, and the antenna efficiency as defined for transmission. If the efficiency is high, *i.e.*, if the antenna resistance is largely radiation resistance, the energy that can be abstracted from a passing radio wave tends to be independent

[1] The distance that the ground currents must be transmitted in this way can be reduced greatly in the case of very low frequency antennas by using multiple ground leads spaced perhaps 1000 ft. apart, giving what is termed a "multiple tuned antenna." See N. Lindenblad and W. W. Brown, Main Considerations in Antenna Design, *Proc. I.R.E.*, vol. 14, p. 291, June, 1926, for further information on ground systems for long-wave antennas.

of size until the antenna dimensions approach a half wave length. This is because the equivalent lumped voltage induced in such an antenna is approximately proportional to size, while the radiation resistance is proportional to the square of the size. Hence with perfect matching (*i.e.*, a load resistance equal to antenna resistance) the abstracted energy will be independent of size. With larger physical dimensions the antenna picks up more energy from passing waves, but also tends to be directional. When the loss resistance is appreciable, there is an advantage in making the antenna reasonably large since the loss resistance normally increases more slowly than the square of the size.

The amount of energy that can be abstracted from a passing radio wave by a given antenna tends to decrease as the frequency becomes greater. This is because the induced voltage is independent of frequency while the radiation resistance increases rapidly with frequency. When the antenna dimensions do not exceed a half wave length, the radiation resistance varies as the square of the frequency, and with negligible loss resistance the energy delivered to a load resistance that equals the radiation resistance will be inversely proportional to the square of the frequency. As a result increasing attention must be paid to the amount of energy abstracted as the frequency becomes high.

A receiving antenna should abstract sufficient energy from passing waves so that the noise level resulting from static and other natural causes under normal receiving conditions is at least comparable with the thermal-agitation energy existing in the input of the receiver. The signal-to-noise ratio cannot then be improved by further increase in received energy, and the antenna system is adequate as far as energy pick-up is concerned.

When it is desired to avoid marked directional effects in the reception of signals of broadcast and lower frequencies, a single wire with one end elevated and the other grounded through the receiver is ordinarily used. The height rather than the length of such an antenna determines the induced voltage since waves of broadcast and lower frequencies are vertically polarized in the vicinity of the earth. A height of only a few feet is sufficient if great care is taken to eliminate all loss resistance and to match the receiver input perfectly with the antenna, but it is usually simpler to make the height at least 15 to 20 ft. so that the penalty of not realizing ideal conditions is less.

The usual non-directional antenna for the reception of short-wave signals is a single wire perhaps 25 to 50 ft. long. This may be used as a vertically or horizontally polarized antenna since short-wave signals ordinarily contain both components. However, most man-made disturbances are vertically polarized near the earth, and so can be minimized by a horizontal antenna.

In the special case of automobile-radio antennas, it is customary to use either an antenna in the car top, if this is of wood or fabric, or an arrangement employing metal plates suspended below the running boards, with one receiver input terminal connected to these plates and the other to the car. With such antennas the energy pick-up at best is small, and there is consequently considerable advantage in designing the antenna and input circuit to the receiver so that, when the specified antenna constants are realized, there is an efficient coupling to the grid of the first tube.

All-wave Antenna Systems for Broadcast Receivers.—The advent of the all-wave receiver has led to the development of special antenna systems which give adequate pick-up over the full frequency range with a minimum of directivity, and which do not respond to vertically polarized high-frequency noise voltages such as those produced by electrical appliances. An example of such an antenna system is shown in Fig. 419.[1] This arrangement employs a horizontal doublet antenna having an over-all length slightly less than a half wave length at the highest frequency to be received. This antenna has its own ground and is coupled to a balanced non-resonant line by an antenna-coupling unit. The receiving terminal of the balanced line is coupled to the unbalanced input of the receiver through a "line-to-receiver filter" or transformer which has a balanced primary with an electrostatic shield between primary and secondary.

The antenna-coupling unit is designed so that at the high frequencies the antenna functions as a horizontal doublet coupled to the transmission line through transformers L_1L_2 as shown at Fig. 419*b*, with the coupling network so designed that the impedance match between line and antenna is sufficiently good to permit the energy transfer to approach the maximum possible value. At low frequencies the arrangement functions as a flat-top antenna which is coupled to the line through the transformer L_4L_5 as shown in Fig. 419*c*. The transition between doublet and flat-top action occurs in the vicinity of 5 mc and comes about as a result of the fact that at high frequencies the capacity C_2 is a virtual short circuit and so prevents action as a flat-top antenna, while at low frequencies the transformer L_1L_2 is so inefficient as to prevent action as a doublet antenna. The complete antenna-coupling unit can therefore be considered as a combination of an impedance-matching network and switching network. This switching is necessary because at the lower frequencies the waves near the earth's surface are vertically polarized and so do not induce voltages in a horizontal antenna. The resistance R shown in the antenna

[1] For further information concerning this all-wave antenna system, with particular reference to the principles involved in the coupling networks, see Harold A. Wheeler and Vernon E. Whitman, The Design of Doublet Antenna Systems, *Proc. I.R.E.*, vol. 24, p. 1257, October, 1936.

ground lead is to prevent the possibility of resonances when this lead is long.

The use of a horizontal antenna carefully balanced to ground greatly reduces the amount of man-made noise picked up at the higher frequencies. The suppression of noise can be still further improved by locating the antenna together with its ground in a quiet location even when this requires the antenna to be some distance away from the receiver. Each

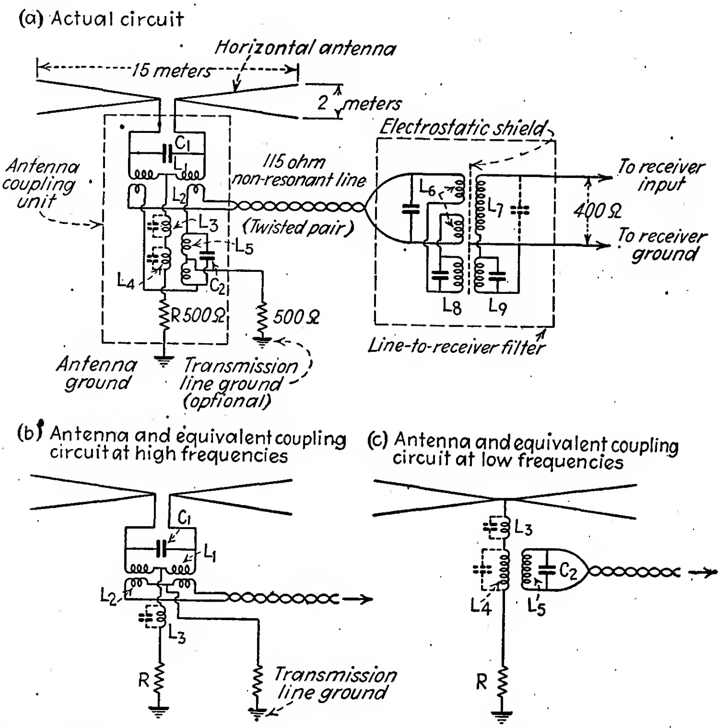


FIG. 419.—All-wave antenna system, together with equivalent circuits at high and low frequencies.

side of the doublet antenna is composed of two wires, as shown in Fig. 419*a*, for the purpose of lowering the reactive component of the antenna impedance in order to simplify the matching of antenna to the transmission line.

The transmission line is ordinarily a twisted pair and terminates in a balanced coupling unit; thus it is relatively free from antenna action. The addition of a transmission-line ground with a 500-ohm resistance as shown still further reduces antenna action by damping out unbalanced currents in the line. The "line-to-receiver" filter consists of a double transformer with condensers, as shown, and with proper design it is possi-

ble to obtain efficient transfer of energy over the entire frequency band from 0.54 to 18 mc, while at the same time making the impedance that the line sees when looking into the filter substantially equal to the characteristic impedance over this entire range, provided the receiver offers a constant input resistance.

The design of the circuits for antenna-coupling and line-to-receiver networks is based upon the theory of wave filters and is beyond the scope of this book. It is sufficient to say that by proper attention to details surprisingly satisfactory results are obtained over the entire frequency band. Thus, in addition to the automatic change in mode of action with frequency, the energy pick-up at high frequencies is within a few decibels of that theoretically obtainable with coupling units individually adjusted to the frequency being received.

Directional Receiving Antennas.—In point-to-point radio communication it is of considerable advantage to employ directional receiving antennas, since in this way interference from static and radio signals coming in from other than the desired direction can be eliminated, with consequent improvement in the signal-to-noise ratio. This is equivalent to increasing the power of the transmitter and, in cases where the bulk of the interference arrives from directions other than that in which the transmitter is located, the benefits from directivity become very great. With a random distribution of interference the gain from a directive receiving antenna is the same as the gain when the same antenna is used for transmitting, while, if most of the energy comes from other than the desired direction, the gain resulting from directivity at the receiver is still greater.

In a directional receiving antenna it is desirable for the directive pattern to have one main lobe with negligible minor lobes, and it is essential that the transmission line be arranged so that no energy reaches the receiver except that picked up by the antenna. In this way strong disturbances from undesired directions deliver practically no energy to the receiver. The most satisfactory directive antenna systems for short-wave work are the fishbone antenna of Fig. 387 and the rhombic antenna of Fig. 393*c*. Both of these are non-resonant and so can operate over a range of frequencies without readjustment. The fishbone antenna is more complicated than the rhomboid, but has the advantage of a directive pattern that is substantially free of minor lobes. Other types of antennas, such as the resonant V, the broadside array, etc., have been used to some extent in reception but are generally less desirable.

The maximum directivity usable in a receiving antenna is set by the angular deviations that can be expected from time to time in the received wave, just as is the case with transmitting antennas. There is also a limit to the physical size of a short-wave directive antenna, since signals received at points separated by distances of the order of 10 wave lengths ordinarily

have more or less random phase difference, and so will not add up properly in the antenna. This causes a loss in directional characteristics since the maintenance of the pattern requires proper phase relations for the different parts of the antenna. It is hence undesirable for the receiving antenna to be more than six to eight wave lengths in size.

At broadcast and lower frequencies the Beverage wave antenna is the most satisfactory directive antenna since it is simple, has excellent directivity, and, as a result of being non-resonant, can be used for the reception of different frequencies without readjustment. In cases where only a single frequency is to be received, as in the transatlantic long-wave telephone, it is also possible to use an array of two to four antennas (such as loops or vertical wires) spaced from half a wave length to two wave lengths apart and coupled together with proper phase relations by use of transmission lines.

In the construction of directional short-wave receiving antennas it is necessary to take special precautions to prevent the directional characteristics from being partially destroyed by the action of near-by conductors, such as towers, guy wires, and power lines. Such objects, if of the right dimensions, will abstract appreciable energy from waves that arrive in unwanted directions and will reradiate this energy to the receiving antenna. If the interference arriving from these unwanted directions is great, this action will largely neutralize the other advantages of the directional antenna. Hence it is customary to locate short-wave directive receiving antennas in spaces that are clear of trees, houses, power and telephone wires, etc., and considerable attention is also paid to minimizing the effect of towers and other metal structures that must necessarily be near the antenna. Wooden telephone poles are particularly satisfactory when they can be used. It is also desirable to locate directive receiving antennas where there will be a minimum of noise of man-made origin. This in general means rural locations removed from main traveled roads, regular air routes, power lines, etc. It is also desirable to bury near-by power wires in underground conduits to prevent the receiving antennas, transmission lines, etc., from picking up disturbances originating on the power system.

Antennas for Ultra-high Frequencies.—With ultra-high frequencies the principal problem involved in non-directional reception is to obtain adequate energy pick-up. Since the energy abstracted by an antenna a half wave length or less in size is very small at such frequencies, it is commonly desirable to employ larger antennas even though this introduces appreciable directivity. It is also essential that the receiving antennas for ultra-high frequencies be located as far as possible above the earth, since, as shown by Eq. (219), the field strength received from a distant transmitter is directly proportional to the antenna height above earth.

Directive antennas used at ultra-high frequencies are commonly of the same types employed at lower frequencies, but with the difference that there is greater flexibility possible in design because of the small physical size represented by a wave length.

Problems

1. A vertical wire 10 meters long carries a current of 5 amp. of frequency 2000 kc. Assuming the wire is in free space, and that the current is uniformly distributed, calculate the strength of the radiated field produced at distances of 1, 10, 50, and 200 km in a direction at right angles to the axis of the wire.

2. In Prob. 1 calculate the rate at which energy flows through a square area $\lambda/4$ on a side when the distances are 1, 10, 50, and 200 km.

3. *a.* A particular antenna consists of a horizontal wire $2\frac{1}{2}$ wave lengths long. Sketch the nature of the current distribution in the antenna, and also sketch the image antenna with its current distribution.

b. Repeat (*a*) for the case where the wire is vertical with its lower end $\lambda/4$ above ground.

4. Discuss the equivalent receiving antenna circuit of Fig. 374 from the point of view of Thévenin's theorem.

5. Discuss qualitatively the reasons that the pattern for the $1\frac{1}{2}$-wave-length wire in Fig. 375 has six lobes (three on each side of the wire).

6. Derive an equation analogous to Eq. (228) but for the case where the wire length is an odd number of half wave lengths.

7. Calculate and plot the directional characteristic in a horizontal plane obtained from two vertical wires spaced one wave length apart and having the same phase.

8. *a.* An antenna system consists of two spaced vertical wires carrying identical currents. Deduce a formula for the directional characteristic in a vertical plane containing the line joining the wires, and for the vertical plane at right angles to this, assuming that the field from the individual radiator is proportional to the cosine of the angle of elevation.

b. Calculate and plot the directive patterns for a spacing of one wave length.

9. Deduce a formula for the directional characteristic in a horizontal plane of three vertical antennas spaced apart a distance d along a line, with phases of $-\alpha$, 0, and α, respectively, and carrying identical currents.

10. In a non-resonant antenna it is possible, by the use of series condensers, to reduce the phase shift of current along the line, even to the point where the current everywhere is of the same phase. Discuss, qualitatively, the effect that this change can be expected to produce in the directional characteristic, with particular reference to the directions of maximum radiation.

11. Discuss the directional characteristic of the array of arrays of Fig. 389 on the basis that this is a two-element array, each element of which is a broadside array, and show that the result is the same as obtained by taking the arrangement to be a broadside array, each element of which is a unidirectional couplet.

12. In a non-resonant V antenna of the type shown at Fig. 393*b*, if the tilt angle is not the optimum value, demonstrate qualitatively that, although the major lobes of radiation from the two sides of the V do not point in the desired direction, the sum of the two lobes does provided the tilt angle is not too far from optimum.

13. Design a horizontal rhomboid antenna that will give optimum directivity at a wave length of 20 meters, when the length of each leg is limited to 60 meters and the radiation is to be concentrated at a vertical angle of 15 deg. above the horizontal. The design includes determination of antenna dimensions (in meters or feet), angles, and height above earth.

14. Calculate and plot the directional characteristic of the antenna of Prob. 13 for the antenna in free space, for the plane of the antenna, and for a vertical plane passing through the terminating resistor and the diagonally opposite apex. From this calculate and plot the directional characteristic in the vertical plane for a perfect earth with the antenna height selected in Prob. 13.

15. Repeat the calculation of directivity in the vertical plane of Prob. 14, assuming an imperfect earth having the reflection coefficients of Fig. 346.

16. Design a horizontal resonant V antenna for service at a wave length of 40 meters, when the length of each leg is 3¾ wave lengths, when the optimum vertical angle is 20 deg., and when a unidirectional characteristic is desired. The design includes specification of antenna dimensions (in meters or feet), apex angle, and height above earth.

17. Calculate and plot the directional pattern for the antenna of Prob. 16 in the horizontal plane, and also in the vertical plane passing through the bisector of the apex angle (assuming a perfect ground).

18. An antenna consists of a horizontal wire a half wave length long and three-fourths of a wave length high. Calculate and plot the directional characteristics in vertical planes that: (*a*) are at right angles to the axis of the wire, (*b*) contain the wire. Make calculations (1) for a perfect earth, (2) for an earth having reflection coefficients as given in Fig. 346.

19. Derive a formula for the directional characteristic in the horizontal plane of an antenna system comprising two vertical wires spaced a distance d apart, carrying currents with magnitudes in the ratio A and having a relative phase Φ.

20. Determine the gain of a vertical grounded antenna that is $5\lambda/8$ high compared with a short vertical wire, using the graphical method.

21. Calculate and plot the resistance component of the impedance of a half-wave horizontal antenna as a function of height above a perfect earth (up to height $= 2\lambda$) by deducing from Fig. 408 the resistance component of the impedance which the *negative* image antenna couples into the actual antenna.

22. Prove that, when n antenna systems are arranged in an array in such a manner that the maximum lobes of radiation all add in phase, then the gain of the combination will be n times the gain of the individual antenna system provided the spacing (or orientation) is such that the mutual impedances between antenna systems are negligible.

23. A parasitic half-wave antenna is $3\lambda/8$ distant from the radiating antenna. Using Fig. 408 determine: (*a*) the magnitude and phase of the induced voltage when the radiating antenna current is I_0; (*b*) the current in the parasitic antenna when it is tuned to resonance; (*c*) the resistance component of the effective impedance of the radiating antenna, taking into account the effect of the parasitic antenna.

24. When a parabolic reflector is used, explain why the focus must be an odd number of quarter wave lengths distant from the nearest part of the reflector.

25. A particular two-wire transmission line is to be constructed from No. 12 wires and must have a characteristic impedance of 600 ohms. Specify the spacing required.

26. In the arrangement shown at Fig. 413*e* explain why the distance m depends upon the antenna height above earth.

27. A half-wave transmitting antenna is to be coupled to a 600-ohm line so that the line will be non-resonant. If the connection to the antenna is made by opening the center of the antenna so that the antenna offers a resistance impedance of 73.4 ohms, design a quarter-wave coupling line of the type shown at Fig. 413*c*.

28. When a particular antenna is connected directly to a transmission line, it is found that the resonances are such that the minimum line current is half the maximum current along the line. Design a stub line that will make the line non-resonant.

29. Describe the mechanism by which unbalances to ground or dissymmetries in a transmission line will cause currents to flow out along the two transmission-line wires in parallel and return via the ground.

30. In (*b*) of Fig. 416*a*, explain: (*a*) the coupling mechanism between line and antenna, (*b*) the reason the line should be an odd number of quarter wave lengths long.

31. Discuss the differences that would exist between short-wave directive antenna systems for communication over a few hundred miles as compared with corresponding antennas intended for communication over great distances.

32. Describe a tuning procedure that would be suitable for placing the antenna system of Fig. 397*c* in operation.

33. Calculate and plot the directivity in a vertical plane of grounded vertical antennas of heights $\lambda/8$, $3\lambda/8$, $\lambda/2$, $5\lambda/8$, and $3\lambda/4$. Plot on the same polar coordinates and reduce all cases to the same radiation along the horizontal. From the results discuss the effect of antenna height from the point of view of broadcast coverage.

34. Design a fishbone antenna to receive long-distance signals in the frequency range 15 to 7.5 mc. The design includes specification of number, length, and spacing of the collector antennas, and the height of the array.

35. Explain why a large receiving antenna is of advantage when the receiver is not particularly sensitive, but is of little or no advantage with a sensitive receiver.

36. In the all-wave receiving antenna of Fig. 419 what would be the practical effect of: (*a*) a poorer impedance match between the antenna and the line, (*b*) a poorer impedance match between the line and the receiver input, (*c*) removal of electrostatic shielding in the line-to-receiver filter, (*d*) increasing the length of the transmission line?

CHAPTER XVI

RADIO AIDS TO NAVIGATION

139. Fundamental Principles of Radio Direction Finding.[1]—The fact that radio waves propagate away from the transmitter along a great-circle route can be utilized in direction-finding work. Thus a ship or airplane can obtain its location by determining the direction of the radio waves sent out by transmitters at known locations. Similarly it is possible to determine the location of a radio transmitter by taking bearings on the radio waves at two receiving locations.

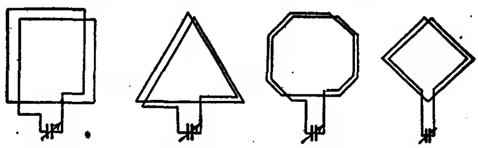


FIG. 420.—Examples of typical loop antennas.

Analysis of Loop Characteristics.—Most practical direction-finding systems make use of a loop antenna, which is essentially a large coil of any conveniently-shaped section (see Fig. 420, for examples). Such an antenna has the directional characteristic shown in Fig. 421 and abstracts energy from passing waves as a result of phase differences between the voltages induced in the opposite legs. Thus consider the case of a rectangular loop in the path of a vertically polarized radio wave. When the plane of the loop is perpendicular to the direction of travel of the waves, the voltages induced in the two legs are of equal magnitude and the same phase, and, being directed around the loop in opposite sense, cancel each other and result in zero response. As the plane of the loop is brought nearer to parallel with the direction of wave travel, the wave front reaches the two legs at slightly different times, causing a phase difference between the voltages induced in the two legs and giving rise to a resultant voltage that acts around the loop and is maximum when the plane of the loop is parallel to the direction along which the waves travel.

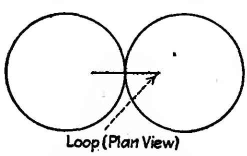


FIG. 421.—Directional characteristic of loop antenna. This applies to loops of all shapes.

[1] For further information on direction finding and for references to the extensive literature on the subject one should read R. L. Smith-Rose, Radio Direction Finding by Transmission and Reception, *Proc. I.R.E.*, vol. 17, p. 425, March, 1929.

Vector diagrams illustrating the situation for several loop orientations are shown in Fig. 422.

The resultant voltage acting around a rectangular loop is given by the following equation:[1]

$$\text{Resultant voltage acting around loop} = 2\epsilon l N \sin\left(\frac{\pi s}{\lambda}\cos\theta\right) \qquad (251a)$$

where

ϵ = strength of radio wave in volts per meter
l = height of loop in meters
s = width of loop in meters
N = number of turns in loop
λ = wave length of radio wave in meters
θ = direction of travel of wave with respect to plane of loop.

In practical loops the size is small compared with a wave length so that $\sin\left(\frac{\pi s}{\lambda}\cos\theta\right)$ may be written as $\frac{\pi s}{\lambda}\cos\theta$ without appreciable error, giving

$$\text{Resultant voltage acting around loop} = 2\pi\epsilon N\frac{ls}{\lambda}\cos\theta \qquad (251b)$$

$$= 2\pi\epsilon N\frac{(\text{loop area})}{\lambda}\cos\theta \qquad (251c)$$

Equation (251c) applies to loops of all shapes provided only that the loop is small compared with a wave length.[2]

The effectiveness of a loop as a means of abstracting energy from passing waves (or for radiating energy) is low because of the tendency of the opposite legs to cancel each other's effects. The practical usefulness of the loop arises from its directional characteristic, its convenient physical form, and its independence of the ground.

Direction Finding with Loop Antennas.—The direction of travel of a radio wave can be determined by using a loop antenna to excite a radio receiver. The loop is rotated until zero response of the receiver indicates that the plane of the loop is perpendicular to the direction in which the

[1] This formula can be readily derived as follows: The voltage induced in each vertical leg is ϵNl, while the phase difference between the voltages is $\frac{2\pi s}{\lambda}\cos\theta$ radians, since the wave front must travel a distance $s\cos\theta$ to pass from one leg to the other. Subtracting the voltages in the two legs while taking into account this phase difference gives Eq. (251).

[2] A discussion of loops not small compared with a wave length is given by L. S. Palmer and L. L. K. Honeyball, The Action of Short-wave Frame Aerials, *Proc. I.R.E.*, vol. 20, p. 1345, August, 1932; V. I. Bashenoff and N. A. Mjasoedoff, The Effective Height of Closed Aerials, *Proc. I.R.E.*, vol. 19, p. 984, June, 1931.

wave travels (see Fig. 421). In using the loop the position for minimum rather than maximum response is used because the percentage of change of response with small change in loop position is much greater in the vicinity of the loop minimum.

A simple loop antenna employed as described gives the bearing angle of the passing radio waves but leaves a 180-deg. uncertainty in the actual direction of the transmitting station. The sense of the bearing can be determined by making use of a vertical antenna in conjunction with the loop. The vector diagram of Fig. 422 shows that the polarity of the resultant voltage acting around the loop depends on the direction from which the waves arrive, so that, if a small amount of pick-up from a vertical antenna is suitably coupled into the loop, the antenna action will cause one lobe of the loop pattern to be enlarged and the other to be diminished, as shown in Fig. 423. When the two forms of pick-up are of the same magnitude, the directional pattern reduces to a cardiod. The procedure for direction finding is hence to obtain the bearing by adjusting the loop for zero response with the vertical antenna disconnected, after

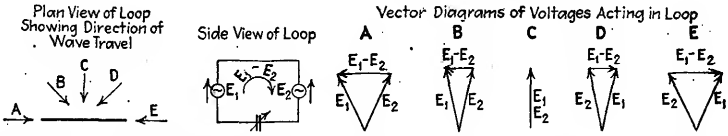


FIG. 422.—Vector diagrams showing how the voltages induced in the two sides of a loop by a passing radio wave combine to give a resultant voltage acting around the loop.

which the 180-deg. uncertainty in the bearing is eliminated by rotating the loop 90-deg. in a specified direction and coupling the vertical antenna to the loop. If the addition of the vertical antenna increases the signal, the sense is one way, while a diminution of signal indicates the opposite sense.

The vertical antenna used to obtain the sense of the bearing can be any convenient arrangement located near the loop, and must be coupled so that the voltage induced in the loop circuit will either add to or subtract from the resultant loop voltage. A practical arrangement is shown in Fig. 423. Here a resistance of several thousand ohms is placed in series with the antenna to make the current in the antenna substantially in phase with the induced voltage and independent of frequency. This current then passes through a coil that is inductively coupled to the loop, thereby causing the vertical antenna voltage acting in the loop to be 90 deg. out of phase with the voltage induced in the vertical antenna. This gives the required phase conditions, since reference to Fig. 422 shows that the resultant voltage acting around the loop is also almost 90 deg. out of phase with the voltages induced in the vertical sides of the loop. In Fig. 423 the required antenna effect is obtained by connecting the

antenna to one end of the antenna coil. The amount of vertical antenna effect must not appreciably exceed the loop pick-up, but at the same time it must be sufficient to produce a noticeable effect upon the directional pattern.

Errors in Loop Bearings and Their Elimination.—The bearings obtained by the use of a loop antenna are accurate only when the loop is remote from

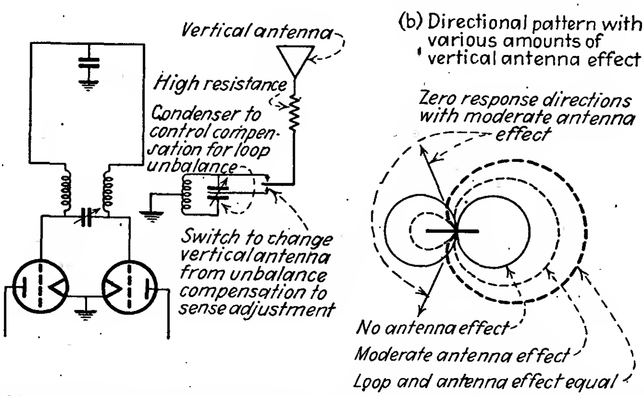


FIG. 423.—Circuit diagram of practical radio compass, together with typical directional patterns that result when loop is combined with the vertical antenna.

metal objects, when it is balanced electrostatically to ground, and when down-coming horizontally polarized waves are absent. Wires and other metal objects abstract energy from passing waves and then produce induction and reradiation fields that induce spurious voltages in a near-by loop. The resulting error in bearing can usually be allowed for by an experimentally determined correction curve, provided the metal objects

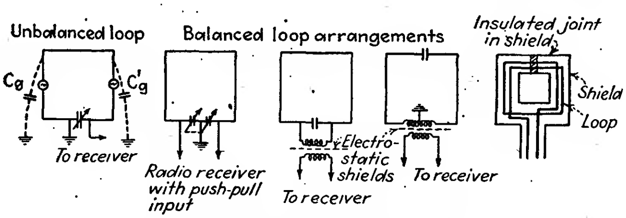


FIG. 424.—Unbalanced, balanced, and shielded loop arrangements.

are always in the same relation to the loop. The correction required will ordinarily depend somewhat upon frequency.

Spurious bearings can also be expected unless the loop is electrostatically balanced with respect to ground. Thus, referring to the left-hand diagram in Fig. 424, after the loop is in the zero signal position the voltages induced in the vertical legs are of the same magnitude and phase, and, being directed around the loop in opposite directions, produce no

resultant voltage around the loop circuit. However, because of the dissymmetry that results from grounding one side of the tuning condenser, capacity currents flowing to ground through C_g' flow through the tuning condenser, whereas the corresponding currents through C_g do not, and this unbalance causes a signal to be delivered to the receiver even though the loop is in the position for zero response. The resulting effect is equivalent to coupling a small amount of vertical antenna pick-up into the loop circuit, and results in a spurious bearing, as seen from the directional characteristic given in Fig. 423*b*. It will be noted from Fig. 423*b* that a small amount of vertical antenna action such as produced as the result of unbalance causes the angle between the two zero positions to differ by less than 180 deg., so that the presence of unbalance in a loop can be tested for by finding a null position and then rotating the loop 180 deg. and noting whether or not a null is again realized. Errors from unbalance can be minimized by using circuit arrangements that are symmetrical with respect to ground, such as shown in Fig. 424. It is also helpful to enclose the loop in an electrostatic shield, such as a metal housing broken by an insulated bushing, as shown in Fig. 424, so that the shield does not act as a short-circuited turn. Such a shield insures that all parts of the loop will always have the same capacity to ground irrespective of the loop orientation or of neighboring objects. Finally, where maximum accuracy is required it is possible to compensate for residual unbalances by connecting the antenna to the center of a small balancing condenser such as shown in Fig. 423, in order to introduce a controllable amount of compensating antenna effect. The proper adjustment of this balancing condenser is obtained experimentally as the setting for which a 180-deg. rotation of the loop does not affect the null.

When horizontally polarized downward-traveling waves are present, the horizontal members of the loop have voltages induced in them that do not give zero resultant voltage when the plane of the loop is perpendicular to the bearing of the radio waves. This causes the minimum signal to occur at a false position, and in some cases makes it impossible to obtain zero response at any loop position. Since horizontally polarized downcoming waves are produced by the action of the ionosphere, the error is negligible near the transmitter where the sky wave is much weaker than the ground wave, but becomes of increasing importance as the distance is increased. When this type of error is present, the bearing of the waves as observed by the loop appears to vary continuously. At broadcast and lower frequencies, where loops are most commonly used, the sky wave is much stronger at night than in the daytime, so that the error from down-coming sky waves is often termed *night effect*. The "night effect" type of error sets a limit to the usefulness of the loop as a direction-finding device, and makes it necessary to replace the loop by an Adcock

aerial, as described below, when it is necessary to take bearings in the presence of strong sky waves.

When the straight-line path between radio transmitter and receiver follows along a coast line, the direction from which the radio waves arrive at the receiver often differs from the true bearing of the transmitter. This comes about as a result of the higher ground-wave attenuation over the land, and makes it necessary to distrust all bearings obtained under this particular situation. Similar behavior is often found in mountainous regions, since the attenuation of a wave traveling up a valley is less than for a wave traveling crosswise from ridge to ridge.

A properly balanced and calibrated loop will ordinarily give bearings on near-by radio stations that are accurate to within ½ to 2 deg. The accuracy diminishes, however, as the distance to the transmitter increases, is greater during the day than at night, is reduced as the frequency is raised, and is greater when the transmission is over water than when over land. This is because all these factors affect the ratio of ground wave to downward-traveling sky wave. The maximum distance for which satisfactory bearings can be obtained in the day-time is of the order of 50 to 200 miles at frequencies of about 500 kc, and may be as great as several thousand miles at very low radio frequencies.

When radio direction finding is employed to enable ships, airplanes, etc., to determine their location, it makes no difference whether the ship observes the bearings of waves radiated from known locations or whether receivers at known positions are used to determine the bearings of waves sent out from the ship. This is because the Rayleigh-Carson reciprocal theorem (see Sec. 130) shows that under the conditions where accurate bearings are possible (no effect from down-coming sky waves) the observed bearings are unaffected by interchanging the transmitter and direction-finding receiver.

Adcock Antenna.[1]—The errors in bearing caused by down-coming horizontally polarized sky waves can be eliminated by replacing the loop antenna with the Adcock aerial system, which in its simplest form consists of two spaced vertical antennas connected as shown in Fig. 425. The action of such an antenna as far as vertically polarized waves are concerned is identical with the loop since the resultant current in the output coil of the Adcock antenna is proportional to the vector difference of the voltages induced in the two vertical members, exactly as is the case with the loop. Horizontally polarized down-coming waves do not affect the Adcock antenna, however, since the voltages induced in the two horizontal

[1] For further discussion of the Adcock antenna, see Smith-Rose, *loc. cit.;* R. A. Watson Watt, Polarization Errors in Direction Finders, *Wireless Eng.*, vol. 13, p. 3, January, 1936; R. H. Barfield, Recent Developments in Direction Finding Apparatus, *Exp. Wireless and Wireless Eng.*, vol. 7, p. 262, May, 1930.

members are the same in magnitude and phase and cancel out as a result of the circuit arrangements. By maintaining symmetry with respect to ground and by enclosing in an electrostatic shield, the Adcock antenna makes it possible to obtain accurate bearings up to several hundred kilometers at frequencies in excess of 5000 kc under conditions where a loop antenna is utterly useless. The practical value of the Adcock antenna is, however, limited by the fact that the energy pick-up is the same as that of a one-turn loop, and so tends to be quite small. Hence the Adcock antenna must be relatively large compared with a loop, and so is not suitable for many applications.

Goniometer Arrangements.—When it is desired to avoid the necessity of rotating a loop or Adcock antenna, it is possible to obtain the effect of rotation by using two loop or Adcock antennas at right angles to each other and combining the outputs in a goniometer. The goniometer con-

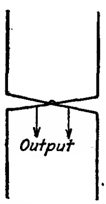


FIG. 425.—Simple form of Adcock aerial system.

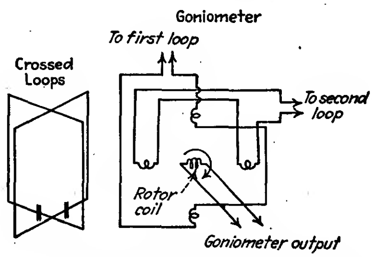


FIG. 426.—Goniometer arrangement for shifting directional characteristic of loop antenna without moving the loop.

sists of two pairs of primary coils (one pair for each antenna) arranged at right angles to each other and coupled to a secondary coil as shown in Fig. 426. If the goniometer is built so that the mutual inductance between each pair of primary coils and the secondary is proportional to the cosine of the angle that the axis of the secondary coil makes with the axis of the primary coils, then for any position of the secondary the directional characteristic of the antenna system will be identical in form with the directional characteristic of a single loop or Adcock antenna. The orientation of this characteristic depends upon the position of the rotating coil, however, and can be changed at will by rotating the goniometer secondary just as the orientation of the directional characteristic of a loop is changed by rotating the loop.

Homing Devices.—A homing device is a form of direction finder which gives a visual or meter indication of the orientation of the direction-finding equipment with respect to the direction of travel of the radio waves. A typical circuit arrangement is shown in Fig. 427. Here loop and vertical antennas are combined to give a cardioid directional pattern,

and the output of the receiver is rectified and passed through a direct-current zero-center galvanometer. The polarity of the loop output is continually reversed by means of a commutator so that the directional pattern is shifted back and forth between the solid and dotted cardioids of Fig. 427. A second commutator synchronous with the first reverses the terminals of the galvanometer in synchronism with the reversals of the loop polarity, so that the galvanometer deflection represents the difference between the receiver output for the two loop polarities. If the loop is perpendicular to the direction in which the signals travel, the antenna-system output is the same for both loop polarities, so that equal d-c currents will pass through the galvanometer alternately in opposite directions and there will be no deflection. However, if the plane of the loop is not perpendicular to the bearing of the signal, the output of the antenna system will depend upon the loop polarity as seen in Fig. 427, and a resultant d-c current will flow through the galvanometer, causing a deflection

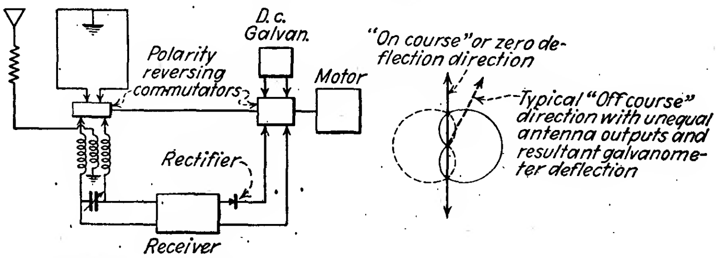


FIG. 427.—Circuit diagram of simple homing device, together with diagrams of the antenna directional patterns for the two loop polarities.

to the right or left according to which side of the zero direction the signals arrive from. The sense of the bearing obtained in this way can be determined from the fact that, if the loop is deflected from the position giving zero output, the direction in which the galvanometer deflects will depend upon the sense of the bearing.

The actual details of the circuit arrangements of homing devices may vary considerably. Thus in some cases the polarity of the vertical antenna instead of the loop is reversed, and in many cases vacuum tubes are used for switching instead of the mechanical arrangement illustrated in Fig. 427.

Homing devices are particularly useful in guiding planes, ships, etc., to a base. Thus an airplane sent out from an aircraft carrier can be guided back to the ship by means of a homing device set to give an "on-course" indication (zero galvanometer deflection) when the plane is headed in the direction of travel of the radio signals sent out from the carrier. It is then merely necessary to make sure that the airplane is headed toward rather than away from the carrier, which is done by intentionally turning

the plane to one side and noting which way the output galvanometer deflects, and then following the radio waves back to their source.

140. The Radio Range.[1]—The radio range is a type of radio beacon which lays down a course in a predetermined direction. The radio ranges in common use are of the aural type and employ two crossed-loop or Adcock antennas which are alternately excited from a common source of radio-frequency power. The directional characteristics of the crossed antennas are shown in Fig. 428, where it is seen that in certain directions from the transmitter the signals from the two antennas are of equal strength. By interlocking the two antennas so that one of them is always radiating energy, and then sending out complementary signals such as N(— -) and A(- —), the signal heard along the equisignal line is a continuous dash, while at points to the side of this equisignal course either one or the other of the code characters dominates, depending on the side

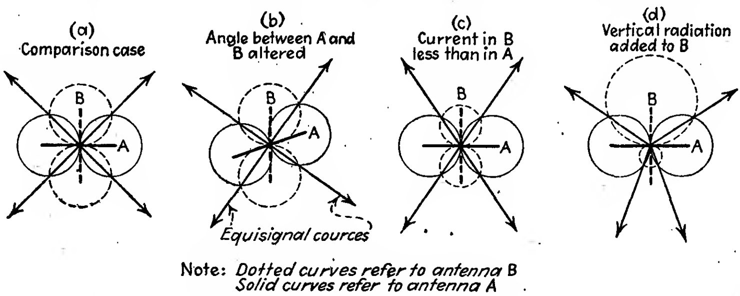


FIG. 428.—Equisignal courses from an aural radio range under various conditions of adjustment.

of the course on which one is located. The principal air routes of the United States are marked by this type of radio range.

The various courses of the radio range can be aligned to the actual routes followed by air travel in several ways. These include varying the angle between the crossed antennas, supplying unequal power to the two antenna systems, and adding a small amount of additional radiation from a vertical antenna, as illustrated in Fig. 428, and various combinations of these expedients. The radio range transmitters normally operate at frequencies between 200 and 300 kc, and preferably make use of two pairs of spaced vertical radiators arranged to form a pair of crossed Adcock antennas.[2] Large loop transmitting antennas are sometimes

[1] A summary (with bibliography) of the work involved in the development of the radio range is given by J. H. Dellinger, H. Diamond, and F. W. Dunmore, Development of the Visual-type Airway Radio Beacon, *Proc. I.R.E.*, vol. 18, p. 796, May, 1930.

[2] For further information on this type of antenna, see H. Diamond, On the Solution of the Problem of Night Effects with the Radio Range Beacon System, *Proc. I.R.E.*, vol. 21, p. 808, June, 1933.

The proper phase relations between the individual antennas in each pair forming the Adcock antenna can be maintained irrespective of slight changes in tuning caused

employed for transmission, but have the disadvantage that such antennas radiate a strong horizontally polarized wave at high vertical angles. At the low frequencies employed for radio-range operation there is very little change in polarization produced by the refraction in the ionosphere so that a loop gives much more "night effect" error than does an Adcock transmitting antenna. A goniometer arrangement is employed to supply power to the two antennas so that the directional pattern (and hence the equisignal courses) can be shifted at will without disturbing the antenna structure.

Another form of radio range that is being used experimentally is termed the *visual* type. It employs the same antenna system as the aural type, but, instead of alternately exciting the separate antennas, it radiates a carrier wave modulated at a convenient low frequency (usually 65 cycles) from one antenna and the same carrier frequency modulated at a different frequency (usually 86.7 cycles) from the second antenna. The result of these two radiations is to produce equisignal zones in which the 65- and 86.7-cycle side bands are of the same amplitude. An airplane making use of signals from the visual-type radio range is supplied with a radio receiver which delivers its output to two reeds tuned to 65 and 86.7 cycles, respectively, which vibrate with an amplitude corresponding to the strength of the corresponding side-band components. These reeds are mounted on the pilot's instrument board with the tips visible. The tips are painted white and when in vibration give white lines proportional to the side bands involved. When on the course laid out by the visual radio range, the two reeds vibrate with equal amplitude; when off the course, one will vibrate with greater amplitude than the other and will indicate in which direction the true course lies.

The antenna for receiving radio-range signals must be vertical with all horizontal components of the antenna structure symmetrically arranged.[1] In this way there will be no response to horizontally polarized down-coming waves such as radiated by a loop transmitting antenna or produced by rotation of the polarization in the ionosphere.

With a suitable receiving antenna and an Adcock transmitting antenna the course indications of a radio range operating at frequencies between 200 and 300 kc can be depended upon to distances in excess of 100 miles for both day and night, while with a loop transmitting antenna the night

---

by weather, etc., using means described by F. G. Kear, Maintaining the Directivity of Antenna Arrays, *Proc. I.R.E.*, vol. 22, p. 847, July, 1934, and Hans Roder, Elimination of Phase Shifts between the Currents in Two Antennas, *Proc. I.R.E.*, vol. 22, p. 374, March, 1934.

[1] For further information see H. Diamond and G. L. Davies, Characteristics of Airplane Antennas for Radio Range Beacon Reception, *Proc. I.R.E.*, vol. 20, p. 346, February, 1932.

range is usually reduced to less than half this distance. The width of the equisignal course depends upon the transmitter adjustments, but is commonly 5 to 10 per cent of the distance to the transmitter.

141. Blind-landing Systems.—A number of radio devices have been devised to permit the blind landing of airplanes. In one arrangement two marker beacons mounted on trucks are driven to points approximately ¼ and 2 miles distant from the airport at locations such as to indicate the proper into-the-wind approach for landing. Before landing the pilot first obtains the position of these marker beacons from the airport radio operator, and then determines his own position with respect to the airport by ascertaining with the aid of his own radio receiver when the plane is directly over the marker beacons. A blind landing is then made using a sensitive altimeter to give the height above ground and depending upon the gyrocompass and other navigating instruments on the plane to control the angle of glide.

In another blind-landing system an ultra-high-frequency horizontally polarized landing beacon having a directional characteristic such as shown in Fig. 429 is used to control the glide.[1] This landing beacon is outlined by marker beacons, and the landing is accomplished by approaching the ultra-high-frequency beacon at a specified elevation. A marker beacon indicates the point at which the glide should begin, and the landing is accomplished by simply following the beam down to earth while keeping the signals received from the landing beacon at a constant level. The equisignal line followed has the general character illustrated in Fig. 429, and by giving the landing beacon the proper directional characteristic the equisignal line can be readily made correct for a perfect landing.

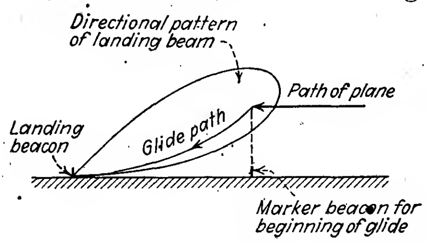


FIG. 429.—Blind-landing system in which the plane glides down a path of constant signal intensity when landing.

Problems

1. Determine the height of a vertical antenna that will have the same induced voltage as the resultant loop voltage of a loop 3 by 2 ft. with 20 turns, when the wave length is 1000 meters.

2. Discuss the factors that set a practical limit to the number of turns that can be used on a given-sized loop frame.

3. Describe the three-dimensional directional pattern of a loop that is in free space (remote from ground).

4. Explain why the zero-signal position of a loop is not affected by horizontally polarized waves traveling parallel to the ground or by vertically polarized down-coming waves, but is affected by horizontally polarized down-coming waves.

[1] See H. Diamond and F. W. Dunmore, A Radio Beacon and Receiving System for Blind Landing of Aircraft, *Proc. I.R.E.*, vol. 19, p. 585, April, 1931.

5. Derive an equation for the directional characteristic of a goniometer arrangement such as Fig. 426, assuming that the mutual inductance between the rotating coil and a pair of stator coils is proportional to the cosine of the included angle.

6. Demonstrate that the Adcock transmitting antenna is superior to the loop for radio-range work only when the ionosphere produces little change in the polarization of the sky wave, and that at high frequencies, where there is a large change in polarization produced by the ionosphere, the Adcock antenna has little if any advantage for transmitting.

CHAPTER XVII

TELEVISION

142. Elements of a System of Television.—A picture can be transmitted electrically by exploring the surface in a systematic manner and transmitting at each instant a current that is proportional to the light intensity of the elementary area being explored at that instant. The result of such a process is to produce a current that varies with time in accordance with the light intensity of successive elements of the picture. This current can be amplified and transmitted directly over wire circuits, or can be modulated upon a carrier and sent out in the form of a radio wave. At the receiving point the received energy is amplified, and in the case of radio reception also rectified, to give at the receiver a current similar to that produced at the transmitter and of adequate amplitude. The picture is then synthesized by employing this received current to control the light intensity of the successive elementary areas corresponding to those at the transmitter. In television this process of picture transmission is speeded up until the effect of a motion picture is obtained.

143. Scanning.—The process of exploring an image to obtain a current that varies with time in accordance with the light intensity of successive areas of the picture is called *scanning*. In all practical systems of television the portion of the picture viewed at any one time is limited by some form of aperture which has a height that is commonly from $\frac{1}{120}$ to $\frac{1}{450}$ of the height of the image and which is commonly square. This aperture is moved relative to the image along successive parallel horizontal paths that are spaced by an amount equal to the height of the aperture, as shown in Fig. 430. With such an arrangement the entire image is systematically covered by means of a series of horizontal lines, the edges of which just barely overlap.

Many methods have been devised for carrying out the scanning operation. Until relatively recently all these were mechanical, and involved such schemes as a flying spot of light controlled by a rotating disk in such a way as to scan the image in the way that the aperture is shown doing in Fig. 430.[1] The practical limitations of such mechanical arrangements

[1] Examples of such mechanical systems are described by Herbert E. Ives, Frank Gray, and M. W. Baldwin, "Image Transmission System for Two-way Television," *Bell System Tech. Jour.*, vol. 9, p. 448, July, 1930. R. D. Kell, "Description of Experimental Television Transmitting Apparatus," *Proc. I.R.E.*, vol. 21, p. 1674, Dec. 1933.

have led to the development of electrically controlled scanning systems which have no moving parts. There are two successful methods of this type suitable for general television work, the "Image Dissector" developed by Farnsworth and the "Iconoscope" of Zworykin.

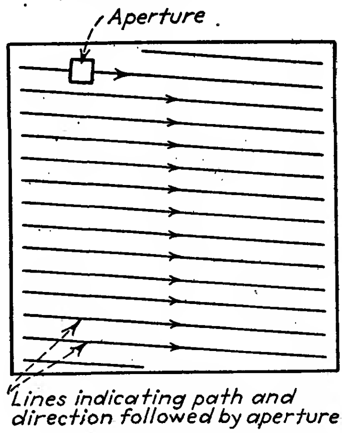


FIG. 430.—Lines indicating path followed by aperture in scanning a picture.

The Farnsworth Image Dissector.[1]—The Farnsworth image dissector was the first successful non-mechanical system of scanning to be placed in operation. The principal features of this device are indicated in Fig. 431. Here C is a translucent cathode, the surface of which is coated with photo-sensitive material, and upon which is focused the optical image of the scene to be transmitted. Each elementary area of the cathode surface accordingly gives a photoelectric emission of electrons proportional to the light intensity of that particular part of the picture. These emitted electrons are then attracted to the anode A, which is at a positive potential with respect to the cathode. The tube is surrounded by a coil carrying a direct current that produces a substantially uniform axial magnetic field. When the strength of this magnetic field is adjusted to the proper value in relation to the anode voltage and tube dimensions,

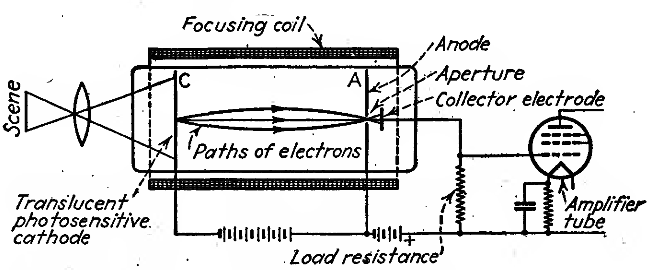


FIG. 431.—Essential features of the Farnsworth image dissector.

electrons emitted from a particular point on the cathode will strike the same spot on the anode surface as shown in Fig. 431 in spite of the fact that many of the photoelectrons emitted from the spot on the cathode have

[1] For further information on the Farnsworth system of television see P. T. Farnsworth, Television by Electron Image Scanning, *Jour. Franklin Inst.*, vol. 218, p. 411, October, 1934; A. H. Brolly, Television by Electronic Methods, *Trans. A.I.E.E.*, vol. 53, p. 1153, August, 1934.

an initial velocity component in a radial direction.[1] Under this condition the distribution of electrons at the anode plane corresponds to the distribution of light intensity upon the cathode, thus giving at the anode what might be termed an electron image of the scene being reproduced.

The anode A is provided with a scanning aperture consisting of a small hole. The electrons passing through the aperture are collected by a collector electrode which hence receives an electron current proportional to the light intensity of the corresponding part of the optical image. The picture is then scanned by displacing the electron image at the anode with respect to the aperture so that the part of the image that supplies electrons to the collector electrode is continuously changing in a systematic manner.

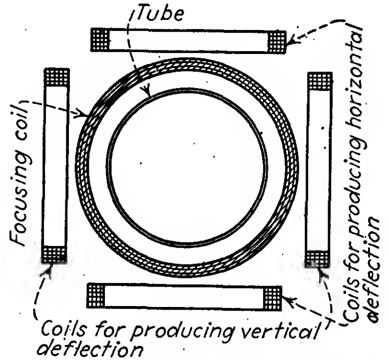


FIG. 432.—Coil system for producing magnetic deflection of electron image and for focusing. The view shown is a cross section through the middle of the tube of Fig. 431 at right angles to the axis.

This displacement of the electron image is accomplished by means of magnetic fields produced by the two pairs of coils shown in Fig. 432. Current passed through the top and bottom pair produces a magnetic field that will deflect the electron image to the right or left, while current passed through the two side coils will in the same manner deflect the electron image up or down. Scanning is accomplished by applying to the side coils a saw-tooth wave such as illustrated in Fig. 433, having a frequency equal to the number of times per second the scene is to be scanned, while the second pair of coils is supplied with a similar saw-tooth wave having a frequency equal to the number of lines per picture times the number of pictures per second. The result is to move the electron image with respect to the aperture in the manner indicated in Fig. 430.

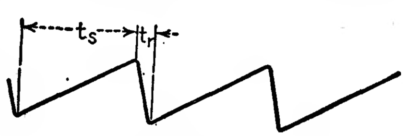


FIG. 433.—Saw-tooth wave used for scanning purposes.

[1] This focusing action comes about as a result of the fact that, while the axial magnetic field exerts no force on an electron traveling parallel to the flux lines (*i.e.*, traveling directly toward the anode), it causes those electrons which have a radial component of velocity to follow a helical path and ultimately to intersect the direct path to the anode. It can be demonstrated that all electrons emitted from the same spot converge at the same point irrespective of the initial radial velocity. This focusing action of an axial magnetic field upon a cathode beam is well known and is discussed many places in the literature. For example see Brolly, *loc. cit.*

Electron Multiplier.—The chief drawback of the pick-up arrangement shown in Fig. 431 is that relatively few electrons pass through the aperture if this is small enough to provide a large number of lines per picture. It has been calculated that under ordinary conditions a 240-line picture of a bright outdoor scene gives only a few hundred electrons when an area equal to the aperture area is scanned. This is so small that, when the output current is amplified, thermal-agitation voltages in the input circuit of the first tube will be comparable to the signal.

Farnsworth has overcome this limitation by making use of secondary emission to multiply the number of electrons that the scanning tube develops. A schematic circuit diagram of one form of electron multiplier developed for this purpose is shown in Fig. 434. Here electrodes B and B' are coated with material that gives prolific secondary emission, and

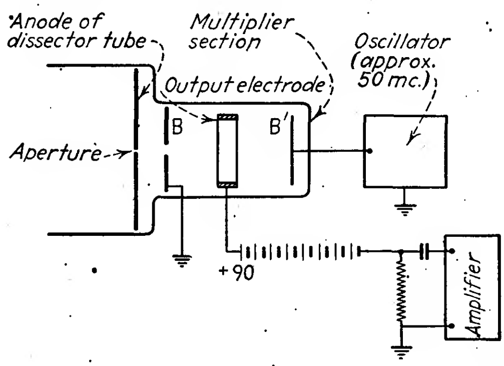


FIG. 434.—Image dissector provided with electron multiplier.

a radio-frequency potential of the order of 50 mc is applied between these electrodes. A central ring anode operated at a moderate positive potential is also provided as shown. Electrons that pass through the aperture and enter the multiplier chamber at the instant when the radio-frequency oscillator is just passing through zero and going positive are attracted to B' and upon striking this electrode emit numerous secondary electrons. If the oscillator frequency is such that the time required for this electron to pass from B to B' corresponds to exactly one-half cycle, then these secondary electrons are emitted at the instant when electrode B is just turning positive with respect to B', so that the secondary electrons are attracted to B with velocity sufficient to produce more secondary electrons, which are then attracted to B', and so on. In this way a very few electrons initially passing through the aperture will in a relatively few round trips be multiplied enormously. A portion of the electrons thus produced is collected by the ring anode and represents the useful output of the multiplier. By proper circuit proportions and operating conditions it is possible to

obtain a stable output that is proportional to the number of primary electrons originally passing through the aperture and is at least 1000 times as great. It is to be noted that the input voltage to the first amplifier tube is now relatively large, so that thermal-agitation effects no longer give trouble. The main source of noise then becomes the shot effect of the original electron flow through the aperture.[1]

The Zworykin Iconoscope.[2]—The iconoscope method of scanning can be explained by reference to Fig. 435. The optical image of the scene under consideration is focused upon the screen P, consisting of a mica or other insulating plate, upon the illuminated side of which is a mosaic surface composed of minute isolated globules that are photosensitive. The manufacturing process used in producing this surface must be such that the individual globules are actually insulated from each other. The back side of the mica or insulating plate has a metal coating that serves

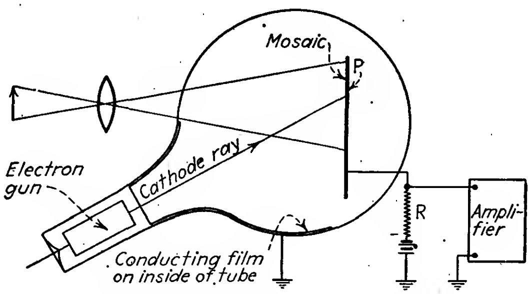


FIG. 435.—Essential features of the Zworykin iconoscope.

as the output electrode, as shown. The optical image focused upon the mosaic photosensitive surface is scanned by means of a cathode-ray beam which is produced by an electron gun similar to that used in ordinary cathode-ray tubes. This cathode-ray beam scans the mosaic surface by means of vertical and horizontal deflecting fields actuated by saw-toothed waves similar to those used in the dissector described above.

The operation of the iconoscope depends upon the fact that each globule of the mosaic surface forms a small condenser with respect to the back plate. When light falls upon a globule, electrons are lost through photoelectric emission, and the globule becomes positively charged with respect to the back plate. This positive charge builds up at a rate proportional to the light intensity. When the cathode-ray beam strikes

[1] Another type of electron multiplier which requires a more complicated tube structure, but which avoids the necessity of an oscillator, is described by V. K. Zworykin, The Secondary Emission Multiplier—A New Electronic Device, *Proc. I.R.E.*, vol. 24, p. 351, March, 1936.

[2] For further description of the iconoscope see V. K. Zworykin, The Iconoscope—A Modern Version of the Electric Eye, *Proc. I.R.E.*, vol. 22, p. 16, January, 1934.

the globule, the negative charge that has been lost by photoemission is replaced by the electrons in the cathode-ray beam, so that the cathode-ray spot can be said to discharge the globule. At the instant of discharge there is a rush of current through the resistance R equal to the positive charge accumulated upon the globule, and hence proportional to the illumination upon the globule. It is to be noted that the current passing through the output resistance R is relatively large because the number of electrons represented by the discharge is equal to the total photoelectric emission since the previous scanning. This means that the total emission over one picture frame (commonly $\frac{1}{30}$ sec.) is utilized, instead of only the photoelectric emission during the instant the spot is being scanned. Actually this storage action is not 100 per cent efficient, since the globule under consideration may receive electrons from other globules and also may lose current by leakage. The actual output is, however, ordinarily several thousand times as great as that obtained from a simple aperture, as in Fig. 431, before electron multiplication.

144. Reproduction of Television Images at the Receiver.—All modern systems of television employ a cathode-ray tube to reproduce the signals at the receiver.[1] The intensity of the luminous spot is controlled by the strength of the incoming signals, while the position is controlled by saw-tooth deflecting waves similar to those employed for scanning at the transmitter. By synchronizing the receiver deflecting waves with the corresponding waves at the transmitter by means described below, the position of the cathode-ray spot on the frame of the received picture will be in the same relative position as the picture element being scanned at the transmitter. The result is then the synthesis at the receiver of the original image.

Cathode-ray tubes for television are essentially the same as the tubes commonly used for observing wave forms, but are larger in order to give a fair-sized image, and operate at a higher anode voltage in order to produce a picture having a satisfactory level of illumination. It is customary to use electrostatic deflection for line scanning and magnetic deflection for the frame scanning. The light intensity of the reproduced image depends upon the watts dissipated per unit area of the fluorescent screen, and for willemite (the usual screen material) the luminous efficiency is about two-thirds that of a tungsten lamp. The fluorescent screen is usually designed so that some light persists for a brief period after the passage of the cathode ray. For optimum results this persistence should be just less than the time required to scan one frame of the picture.[2]

[1] This cathode-ray tube is often given a special name such as *oscillite*, *kinescope*, etc., when used in television, but it is still only a cathode-ray tube.

[2] For more detailed information concerning the properties of cathode-ray tubes with particular reference to television, see V. K. Zworykin, Description of an Experi-

145. Miscellaneous Considerations. *Production of Saw-toothed Waves.* The saw-tooth scanning waves of Fig. 433 must be substantially linear during the period t_s in order for the scanning to take place properly. During the return period t_r the exact shape is not important, but the ratio of return time to scanning time t_r/t_s should be small in order to avoid lost time. The nature of the saw-tooth generator must also be such that the frequency can be readily controlled by means of pulses injected from an external source.

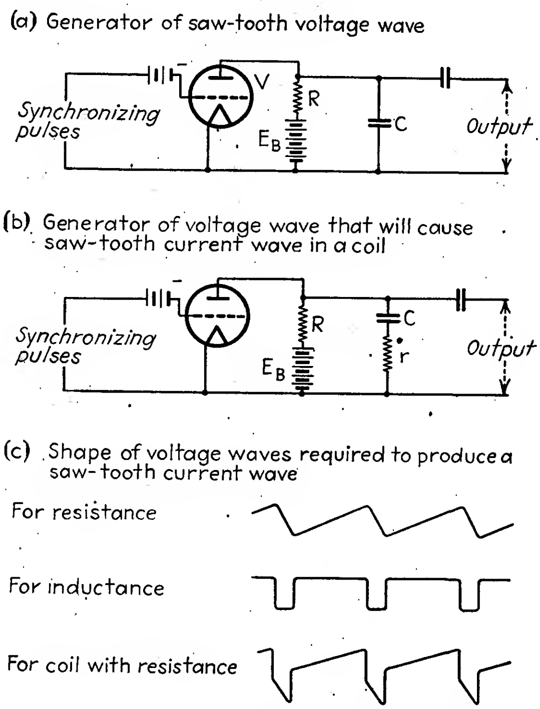


FIG. 436.—Circuit for generating a saw-tooth wave, together with voltage wave shapes required for a saw-tooth current wave through a resistance, an inductance, and a circuit with both resistance and inductance.

With electrostatic deflection a saw-toothed voltage wave is required. This can be generated in a number of ways, a typical arrangement being shown in Fig. 436*a*. Here *V* is a tube that is biased beyond cut-off except during moments when a synchronizing pulse is acting on the grid. The operation can be understood by starting at the moment just after a

mental Television System and the Kinescope, *Proc. I.R.E.*, vol. 21, p. 1655, December, 1933; R. T. Orth, P. A. Richards, and L. B. Headrick, Development of Cathode-ray Tubes for Oscillographic Purposes, *Proc. I.R.E.*, vol. 23, p. 1308, November, 1935; T. B. Perkins and H. W. Kaufman, Luminescent Materials for Cathode-ray Tubes, *Proc. I.R.E.*, vol. 23, p. 1324, November, 1935.

synchronizing pulse has caused the discharge of the condenser C. As soon as the pulse is removed, the tube becomes non-conducting and the condenser starts to charge through the series resistance R. If the circuit proportions are such that the voltage across the condenser is always much smaller than the supply voltage, the resulting condenser voltage will build up linearly, giving a substantially straight-line saw tooth corresponding to the portion t_s in Fig. 433. This process continues until the next synchronizing pulse arrives, when the tube momentarily passes plate current, discharging the condenser C rapidly and giving the return part t_r of the cycle in Fig. 433. After the passing of the synchronizing pulse, the tube again becomes non-conducting and the cycle is repeated. It will be noted that the result is a saw-toothed wave having a frequency controlled by the external synchronizing pulses, and that with proper circuit design a substantially linear wave can be obtained during the scanning portion of the cycle, with a quick return.

The discharge tube in the saw-tooth generator of Fig. 436 can be either a gaseous or vacuum tube when producing low frequencies. However, at higher frequencies, such as used for line scanning, the time lag of gas tubes normally introduces distortion, and a high-vacuum tube is preferable.

When magnetic deflection is employed for scanning, it is necessary to have a saw-toothed wave of current rather than a saw-toothed voltage wave. This complicates matters as the deflecting coils have both inductance and resistance, so that, when a saw-toothed voltage wave is applied, the resulting current is not of the desired wave shape. The simplest manner of handling this situation is to work back from the coil inductance and resistance to determine the wave form of voltage that must be applied to produce the desired shape of current wave. If the coil were a pure inductance, the required shape of the voltage wave would be a simple pulse having a duration equal to the return time, as shown in the second line of Fig. 436c. On the other hand, if the coil were predominantly resistive, the voltage wave would have the same shape as the current wave and so would be saw-toothed, as in the first line of Fig. 436c. In the actual case, where both resistance and inductance are present, it is necessary to combine the wave shapes of resistance and inductance in the proper proportion, such as shown in the last line of Fig. 436c, in order to get a saw-toothed wave of current. Such a voltage wave can be obtained by rearranging the circuit of Fig. 436a as in Fig. 436b. The voltage across the condenser C in this arrangement has the same shape as in Fig. 436a, *i.e.*, is a saw tooth suitable for overcoming the resistance drop of the deflecting coil. The voltage across r, however, is in the form of a pulse, as shown in the second part of Fig. 436c, provided $r \ll R$. This is because, when the condenser discharges, there is a sudden rush of current through

r, while during the charging period the current through r is small and substantially constant.

Synchronization and Synchronizing Pulses.—The preferred method of synchronizing in a television system consists in generating pulses at the transmitter which control the saw-toothed scanning waves used in the pick-up device, and which at the same time are transmitted to the receiver and there used to control the scanning of the reproducing equipment. The required pulses can be generated in a number of ways. One method is to employ a rotating disk provided with holes which allow a light to shine intermittently upon a photoelectric cell. Another method consists in using some form of multivibrator oscillator which is controlled by a vacuum-tube oscillator, the pulses being generated across a small inductance placed somewhere in the multivibrator circuit.

In any case, the end of each line is indicated by a short synchronizing pulse, while the end of each frame is marked by a longer pulse. These

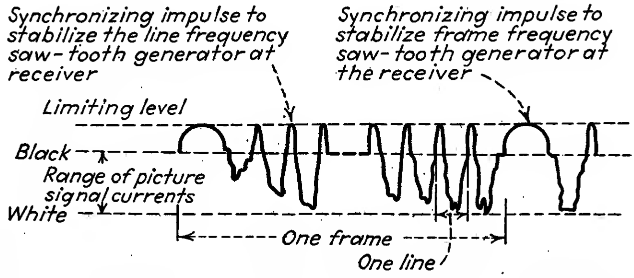


FIG. 437.—Typical television signal showing superimposed synchronizing pulses.

pulses occupy 5 to 15 per cent of the total time, and are separated at the receiver by the difference in duration.

Exact synchronization between transmitter and receiver is obtained by modulating the pulses upon the transmitted carrier, or in the case of wire transmission by transmitting them directly to the receiver along with the signal. In order that the signal currents will not interfere with the synchronizing action, the polarity of the transmitted synchronizing pulses is made to correspond to the polarity of black, and the amplitude of the pulse is made appreciably greater than the amplitude of the signal current for a black portion of the image. In this way the synchronizing pulses are always larger than the maximum possible signal current. In order to prevent overloading of the transmitting equipment, it is common practice to pass the combined signal and synchronizing currents through a limiting amplifier which limits the peak amplitude to perhaps 50 per cent more than black. The resulting waves as transmitted are shown in Fig. 437. The limiting action at the transmitter can be obtained in a variety of ways, a typical arrangement being shown in Fig. 438, where an amplifier

is so operated that the maximum desired amplitude just barely drives the grid to zero bias. Any additional amplitude causes very little additional output because the resulting grid current in flowing through the high resistance in series with the grid prevents the grid voltage from rising appreciably above zero.

By transmitting the synchronizing pulses with a polarity corresponding to that of black in the reproduced picture, and at the same time by making these pulses have a duration equal to or slightly greater than the time required for the saw-tooth scanning wave to return from the end to

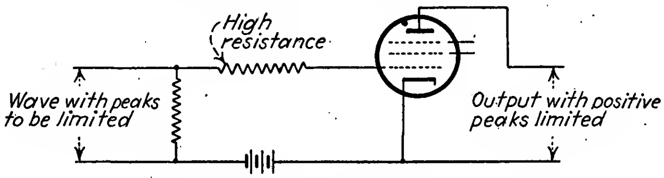


FIG. 438.—Circuit for producing limiting action of television signal peaks.

the beginning of its linear portion, the cathode-ray spot is blanked out during the return period and no disturbance is produced in the picture.

At the receiver the synchronizing pulses are separated from the signal currents by taking advantage of the difference in amplitude. This can be conveniently done by biasing an amplifier tube sufficiently far beyond cut-off so that only the synchronizing pulses have sufficient amplitude to cause plate current to flow. The resulting pulses free of signal are then applied to a network such as shown at Fig. 439, which separates the line and frame pulses by utilizing their difference in duration and frequency.

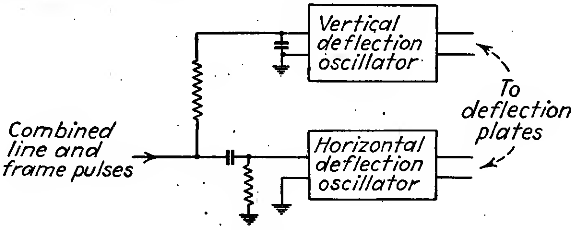


FIG. 439.—Network for separating line and frame synchronizing pulses.

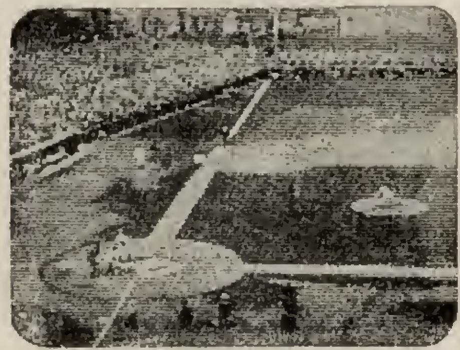
Frequency Band and Picture Detail.—The detail of a television picture is determined by the number of scanning lines. This is obviously true for gradations in the light intensity of the image in the vertical direction, because any change taking place in a distance that is less than the width of a line obviously cannot be reproduced. The situation with respect to the horizontal direction is somewhat more complicated but comes to about the same result.

Studies show that it is necessary to have at least 180 lines to reproduce a scene fairly satisfactorily, that 240 lines give a fair picture, and that 480

lines are practically equivalent to a home movie. At the present time it is considered that commercial television must have at least 240 lines, and there is a tendency to standardize on 350 to 450 lines.[1] The effect of the number of lines upon the detail is shown in Fig. 440.

The optimum viewing position for a television image is such that the distance between the centers of adjacent lines subtends approximately 2′ of arc. This insures that the lines will merge together and give a

(a) 120 scanning lines



(b) 240 scanning lines

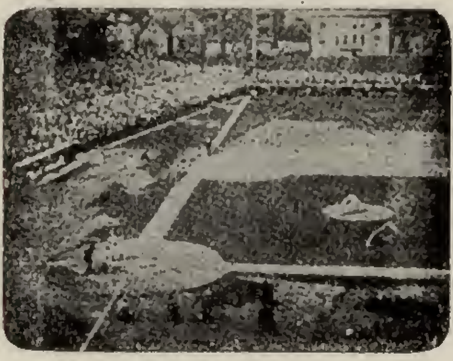


FIG. 440.—Typical television images showing effect of the number of scanning lines.

uniform grainless picture. Closer viewing causes the line structure to stand out, while viewing from a greater distance reduces the amount of detail that can be perceived to less than that actually present in the picture. It is also desirable to make the viewing distance from four to eight times the height of the picture. Hence, if the reproduced image is 12 in. high, the optimum viewing distance is about 6 ft. At this distance

[1] William H. Wenstrom, Notes on Television Definition, *Proc. I.R.E.*, vol. 21, p. 1317, September, 1933; E. W. Engstrom, A Study of Television Image Characteristics, *Proc. I.R.E.*, vol. 21, p. 1631, December, 1933.

it is necessary to have approximately 25 lines per inch for each line to cover 2′ of arc at the observer's eye, so that the picture should be at least 300 lines. Such a calculation is approximate, but gives an idea as to the factors involved in the reproduction of television images.[1]

The frequency band required to transmit a picture having appreciable detail is very great. The worst case is where the image to be transmitted is a checkerboard pattern of alternate dark and light sections as shown in Fig. 441a, with the side of each square equal to the width of the scanning line. In scanning such a picture the output current of the pick-up device will be alternately large and small, so that one cycle of output current corresponds to traversing two squares of the pattern. If the number of lines in the picture is a and the ratio of width to height is R (= aspect ratio), then the total number of squares is a^2R, so that for n pictures per

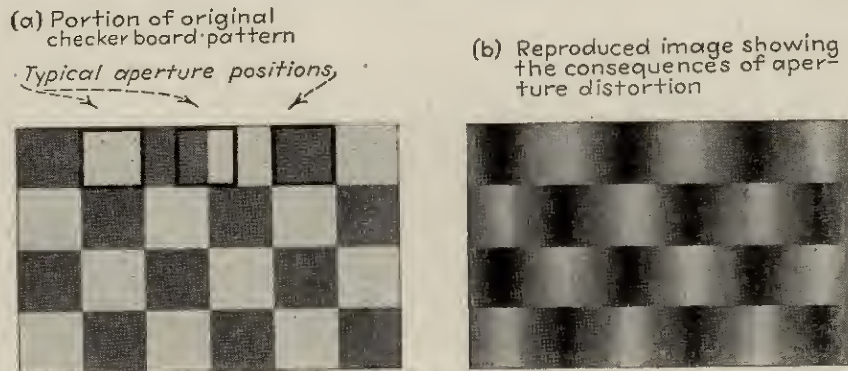


Fig. 441.—Checkerboard image as scanned at the transmitter, and distorted image received as a result of aperture distortion.

second the frequency band required would be $a^2nR/2$ cycles. With ordinary scenes it is found that this gives more detail in the direction of the scanning lines than in the vertical direction. Experience shows the detail is equal in the two directions when the frequency band is only about 70 per cent of the full frequency band, so

$$\left.\begin{array}{l}\text{Actual frequency band}\\ \text{for aspect ratio of } 4/3\end{array}\right\} = 0.35a^2nR \qquad (252)$$

The resulting frequency band required to transmit a television image having good definition is considerable. Thus a 343-line picture, repeated 30 times per second and having an aspect ratio of 4/3, requires a band at least 1,640,000 cycles wide. If 10 per cent is allowed for the synchronizing pulses, the actual frequency band that must be provided is 1,830,000. When this is modulated upon a carrier, the total width of the two side bands is thus nearly 4,000,000 cycles.

[1] Engstrom, *loc. cit.*

As a result of these side-band requirements it is apparent that television pictures can be transmitted by wire only over special circuits. When radio transmission is to be employed, it is necessary that the carrier frequency be extremely high, such as 40 mc or more, so that the side-band width will not be too high a percentage of the carrier frequency. Under such conditions properly designed tuned circuits will handle the modulated television signal satisfactorily without excessive side-band trimming. The use of ultra-high frequencies for the transmission of television signals also has the advantage that there is an enormous frequency band available, and that carrier frequencies can be duplicated at points about 200 miles apart without any interference trouble because of the limited range of propagation.

Aperture Distortion.—The finite size of the aperture at the transmitter introduces distortion by making it impossible to transmit details finer than the area that the aperture represents.[1] This can be made clear by considering what happens when a checkerboard area such as Fig. 441*a* is scanned. As the aperture travels from left to right, it is seen that the average light intensity of the area enclosed by the aperture varies gradually instead of suddenly from light to dark, thus reproducing a distorted pattern as shown at Fig. 441*b*. This effect is equivalent to discriminating against the higher frequency components in the output of the signal.

This situation can be improved to some extent by the use of correcting networks. In this way it is possible to double the rate at which the transition between dark and light takes place. Inasmuch as aperture distortion takes place at both the transmitting and receiving points, it is hence seen that, if correcting networks are employed at both transmitter and receiver, the total amount of aperture distortion is approximately the distortion introduced by a single uncorrected aperture.

Video-frequency Amplifier.—The signal currents obtained from the pick-up tube of the television transmitter are commonly termed video-frequency currents to distinguish them from audio- and radio-frequency currents. Video-frequency amplifiers for increasing the output of a pick-up tube to a level suitable for modulating a radio transmitter, and for amplifying the rectified received signal up to an amplitude suitable for operating the cathode-ray tube, must handle a very wide frequency range with negligible amplitude and phase distortion. The response at low frequencies must extend well below the frequency corresponding to the number of pictures per second, while the highest frequency is given by Eq. (252). Over this range the response should be substantially uniform and should be accompanied by negligible phase

[1] In the case of the iconoscope, the size of the cathode-ray spot scanning the mosaic plate is the effective size of the aperture.

distortion. Phase distortion is particularly important because it is equivalent to different velocities of transmission for different frequencies, and so it obviously distorts the picture by causing certain parts of the detail to arrive either too early or too late. This is in contrast with audio-frequency work, where because of the characteristics of the ear the amount of phase distortion produced by ordinary amplifier circuits has negligible effect.

The usual method of meeting the amplifier requirements is to use a resistance-inductance coupling arrangement of the type discussed in Sec. 46, with a low coupling resistance and the minimum possible shunting capacity. Such an amplifier when properly designed can be made to give substantially uniform amplification with negligible phase distortion up to frequencies that would correspond to the 70 per cent point if the coupling inductance were omitted. The gain per stage is relatively low because of the low coupling resistance that must be used, but is still appreciable.

The amplification of the low frequencies introduces no special problems since amplifiers can be readily built to amplify all frequencies above a few cycles with negligible amplitude and phase distortion. In some cases, however, low-frequency equalization is employed to prevent the dropping off at low frequencies that would take place if the coupling condenser were only moderately large.

Flicker, Power-line Ripple, and Interlaced Scanning.[1]—If the rate of repetition of the television image is not sufficiently high, there will be a pronounced flicker even though the rate of repetition may be sufficient to convey movement satisfactorily. Studies indicate that the lowest repetition rate at which the flicker is not objectionable is approximately 48 times per second, which is considerably greater than the rate required to convey motion.[2]

In selecting the frame frequency it is also necessary to consider the effect of stray power-line currents in the deflecting and anode circuits. If the frame frequency is not an exact submultiple of the power frequency, any stray current of power frequency in the deflecting, anode, or control circuits will cause ripples of various sorts to travel across the field with very annoying psychological effects. If the frame frequency is a submultiple of the power-line frequency, such as 15, 20, 30, or 60 for a

[1] For more detailed information upon these subjects, see E. W. Engstrom, A Study of Television Image Characteristics, Part II, Determination of Frame Frequency for Television in Terms of Flicker Characteristics, *Proc. I.R.E.*, vol. 23, p. 295, April, 1935; R. D. Kell, A. V. Bedford, and M. A. Trainer, Scanning Sequence and Repetition Rate of Television Images, *Proc. I.R.E.*, vol. 24, p. 559, April, 1936.

[2] In motion-picture work where there are 24 frames per second, flicker is avoided by cutting off the light momentarily during the middle period of each frame so that the light flashes on the screen twice for each picture, or 48 times per second.

60-cycle power system, these disturbances, while still present, are now stationary and so are much less noticeable.

Since the usual power-line frequency is 60 cycles and since 60 has no submultiple between 30 and 60, it is therefore necessary that the repetition rate of the television picture be 60 times per second in order to avoid excessive flicker and power-line ripple. This is a relatively high frame frequency, at least three times that required to convey continuous motion, and so is undesirable because of the excessively wide frequency band that results. A fairly satisfactory solution for this situation is to scan the image 60 times per second, but to scan only alternate lines during one scanning period. Thus with a 240-line picture one would scan the odd-numbered lines during the first $\frac{1}{60}$ sec., and then scan the even-numbered lines during the next $\frac{1}{60}$ sec. As far as flicker is concerned the result is equivalent to scanning at the rate of 60 times per second, while from the point of view of detail and frequency band required the result is substantially equivalent to scanning a 240-line picture 30 times a second. An arrangement of this sort is known as *interlaced scanning* and is becoming recognized as the best compromise between low flicker and the narrowest possible frequency band. The only disadvantage is the relatively complicated synchronizing system that is required to interlace the lines accurately.[1]

146. Typical Complete System of Television.—A television transmitter consists of a pick-up tube (iconoscope or image dissector), a saw-tooth wave generator with frequency controlled from a pulse generator, a video-frequency amplifier, and an ultra-high-frequency radio transmitter upon which the output of the video-frequency amplifier is modulated. The amplifier circuits, both video and radio frequency, must be arranged to transmit the wide frequency bands involved without appreciable amplitude or phase distortion. In many respects this layout is similar to that of an ordinary broadcast transmitter, with the pick-up tube taking the place of the microphone and the video-frequency amplifier corresponding to the broadcast audio-frequency system. A block diagram of such a transmitter is shown in Fig. 442*a*.[2]

A typical television receiving system is shown in Fig. 442*b*.[3] Considering for the moment only the picture channel, this consists of a superheterodyne receiver with a high intermediate frequency (5 to 10 mc), with

[1] For a description of suitable synchronizing systems, see Kell, Bedford, and Trainer, *loc. cit.*

[2] For more complete details of a complete television transmitter, see R. D. Kell, A. V. Bedford, and M. A. Trainer, An Experimental Television System—the Transmitter, *Proc. I.R.E.*, vol. 22, p. 1246, November, 1934.

[3] R. S. Holmes, W. L. Carlson, and W. A. Tolson, An Experimental Television System—the Receivers, *Proc. I.R.E.*, vol. 22, p. 1266. November, 1934.

band-pass coupling circuits designed to give a flat response over the necessary frequency range. The intermediate-frequency amplifier is followed by the usual second detector, and the resulting video-frequency

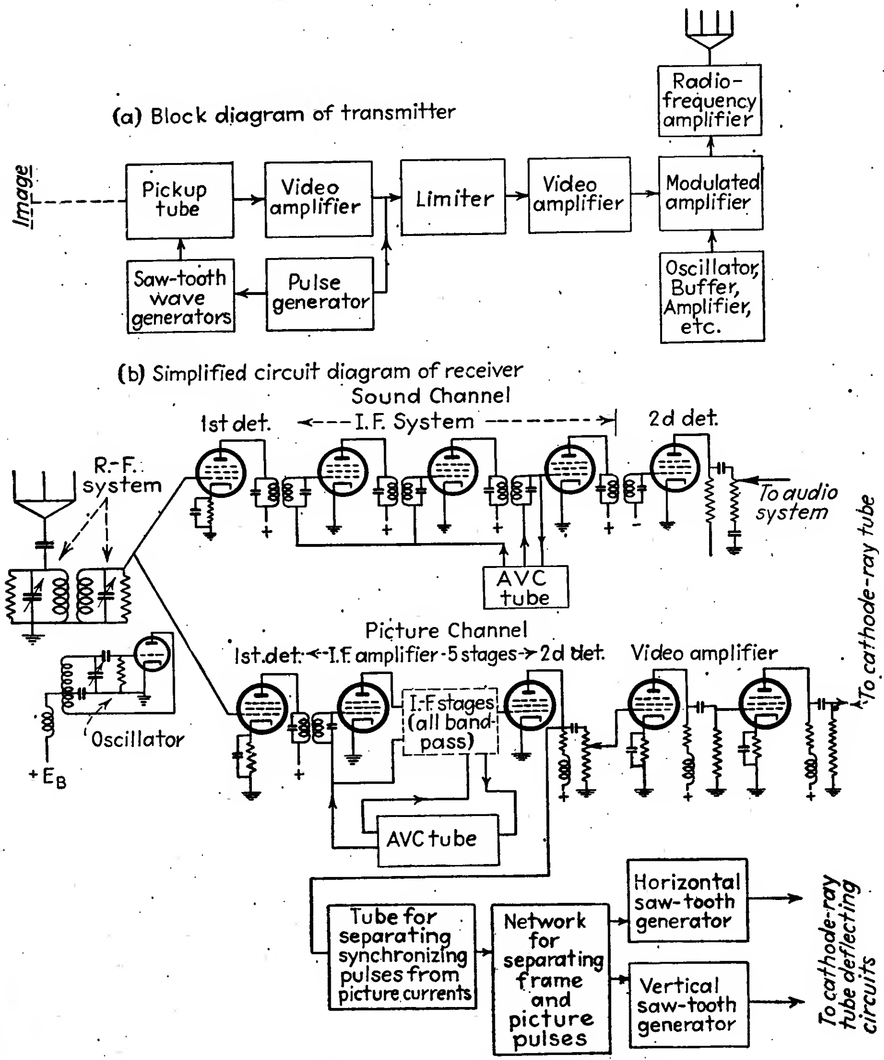


FIG. 442.—Diagram of complete television system.

currents, together with the accompanying synchronizing pulses, are then amplified by the video amplifier to a level suitable for controlling the cathode-ray reproducing tube. Synchronizing pulses free of signal are obtained by also applying the output of the second detector to an ampli-

fier biased so that only these pulses have enough amplitude to develop output. The pulses corresponding to the line frequency are then separated from the frame frequency as shown in Fig. 439 and used to control the frequency of local saw-tooth generators which produce the currents for deflecting the spot of the cathode-ray reproducing tube.

The sound accompaniment for a television picture is most conveniently transmitted by modulating an auxiliary carrier frequency that is just outside the side-band range of the television signals. In the receiver, the antenna circuits and the local oscillator can then be common to both sound and picture channels, but the sound channel has a separate first detector and intermediate-frequency amplifier, as shown in Fig. 442*b*. The intermediate-frequency amplifier for the sound channel should have a relatively narrow response band centered about a frequency differing from the mid-frequency of the picture channel by the difference between sound and picture carriers. The local oscillator then automatically produces the correct beat frequency for the picture channel when the adjustment is such as to bring in the sound channel, and, since the sound channel has a relatively sharp response, it serves as an accurate tuning indicator.

Problems

1. In the Farnsworth image dissector explain the effect upon the reproduced picture of improper strength of the axial magnetic focusing field.

2. In the electron multiplier of Fig. 434, explain why the multiplying action is obtained only for certain oscillator frequencies, which depend upon the multiplier dimensions and the potential of the output electrode.

3. Discuss the effect of leakage between adjacent globules in the mosaic plate of the iconoscope.

4. If the storage process of the iconoscope were 100 per cent efficient, derive a formula giving the improvement in sensitivity over the output obtained by a simple aperture (as the image dissector without electron multiplication).

5. Explain why an ordinary sine wave could not be used for scanning instead of the saw-tooth waves always employed.

6. In the reproduction of television pictures with a cathode-ray tube what would be the effect of using a fluorescent screen material that had: (*a*) a very long persistence, and (*b*) a very short persistence?

7. In the saw-tooth-wave generator of Fig. 436*a* it is necessary that the adjustments be such that the maximum voltage built up across C be limited to a small fraction of the supply voltage E_B. Explain what would happen if this were not the case.

8. Explain why it is not permissible to send the synchronizing pulses with a polarity corresponding to white.

9. Calculate the frequency band required to transmit pictures having 120, 240, and 480 lines, with a frame frequency of 30 and an aspect ratio of $4/3$, when the video current is modulated upon a radio-frequency carrier.

10. Sketch the shape of the output wave of the pick-up tube when one line of the pattern of Fig. 441*a* is scanned for (*a*) no aperture distortion, (*b*) aperture distortion.

11. Design an amplifier to produce a voltage gain of 1000 and capable of handling the output of a scanning tube for a 343-line picture having a frame frequency of 30.

12. When portraying very rapid motion, an interlaced scanning system introduces a peculiar kind of distortion not present with a simple scanning system. Describe the effect.

13. Discuss the number and nature of the controls that are required in a television receiver.

CHAPTER XVIII

SOUND AND SOUND EQUIPMENT

147. Characteristics of Audible Sounds.—Sound is a mechanical vibration lying within the frequency range to which the ear responds, and it is characterized by pitch, loudness, and phase. The pitch represents the frequency and is expressed in cycles per second.[1] The loudness depends upon the amplitude of the wave and is measured either in terms of the pressure in bars (dynes per square centimeter) that is produced by the sound waves, or as the ratio of energy contained in the actual sound to the energy contained in the weakest audible wave of the same frequency. Sound travels in air at a velocity of approximately 1130 ft. per second.

Most actual sounds are complex waves containing several components having frequencies that are in harmonic relation to each other. The fundamental frequency corresponding to these harmonics (or overtones) is called the *pitch* of the complex sound, while the presence of the harmonics determines what is commonly termed the *quality* of the sound. Changing the relative amplitudes of the different components without changing their frequencies hence merely changes the quality without affecting the pitch.

Speech.[2]—The sounds encountered in speech lie in the frequency range 100 to 10,000 cycles. The pitch of speech is determined by the fundamental frequency of the vocal cords, and is about 125 cycles for a normal male voice and about twice as high for a normal female voice. The pitch varies somewhat during speech and tends to rise as the voice intensity is increased. This latter fact makes it possible to estimate the original level of speech independently of the loudness of the sound reaching an observer.

The various speech sounds differ in the way in which the energy is distributed with respect to frequency, in the manner of production, and in the way the sounds are begun and ended. Thus the vowels are relatively powerful sustained sounds produced by using the lungs as bellows and setting the resulting air stream in vibration by means of the vocal cords. The sound so produced is rich in harmonics. The various vowel

[1] This is only approximately the case since recent studies indicate that the pitch of a sound, as judged by the ear, depends to some extent upon the intensity and the overtones present. See Harvey Fletcher, Loudness, Pitch, and Timbre of Musical Tones and Their Relation to the Intensity, the Frequency, and the Overtone Structure, *Jour. Acous. Soc. Amer.*, October, 1934.

[2] For a more detailed discussion of the nature of speech sounds, the reader should consult Harvey Fletcher, "Speech and Hearing," D. Van Nostrand Company, Inc.

sounds are produced by positioning the lips, tongue, etc., so that the throat, mouth, and nasal cavities reinforce harmonics lying within certain frequency ranges as a result of resonance, while tending to suppress other harmonics. Other sounds, such as *w*, *y*, and *h* represent particular ways of beginning the vowel sound, while still others such as *ch*, *sh*, and *f* do not involve the vocal cords.

The power of average conversational speech varies widely with different persons, but under typical conditions for American speech averages approximately 10 μw. If the silent intervals during conversation are excluded, the average is increased to approximately 15 μw. This represents an extremely small amount of energy, as is apparent when it is realized that the total average power that would be produced by all inhabitants of the earth talking simultaneously is less than the power

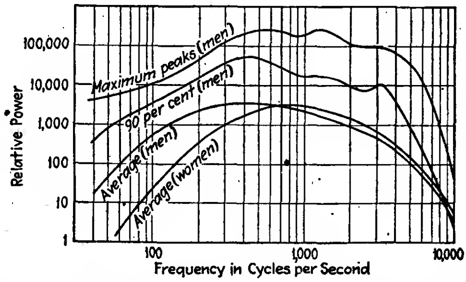


FIG. 443.—Distribution of average and peak speech energy over the frequency spectrum. When the speech is divided into time intervals of ⅛-sec. duration, the peak intensity in 90 per cent of these intervals does not exceed the value given by the curve marked "90 per cent."

radiated from a large broadcasting station. When one shouts as loudly as possible, the average speech power rises to about one hundred times the normal average, while in speaking with as weak a voice as possible without whispering the average power is reduced to approximately one-hundredth. The peak power of the loudest sound encountered in conversation is in the order of 5000 μw, while the power of a faint whisper is in the order of 0.01 μw. This represents an intensity range of 500,000 to 1. The pure vowels and the diphthongs are the most powerful sounds, while the *th*, as in *thin*, is the weakest.

The distribution of speech power over the audible frequency range is approximately as shown in Fig. 443. In the case of male voices the most powerful sounds are more or less evenly distributed throughout the frequency range 500 to 1500 cycles, while the average power is greatest at approximately 500 cycles. The higher frequency peaks, while about as intense as the lower frequency peaks, are less numerous. This is apparent from the 90 per cent curve of Fig. 443, which indicates that, when the speech is divided into intervals of ⅛ sec., the peak intensity

in 90 per cent of the intervals is less than the value given by the curve. In the case of women's voices the higher fundamental pitch tends to shift the energy distribution toward higher frequencies.

A general picture of the frequencies and intensities involved in the different speech sounds as used in normal conversation is shown in Fig. 444. Sounds that have several important frequency ranges appear more than once in the diagram. Powerful sounds are at the top of the figure, weak sounds at the bottom, while positions to the right or left represent important high- or low-frequency components, respectively.

Music.—The frequency range of musical sounds is much greater than with speech. Thus, the bass tuba, bass viol, piano, organ, and drums have fundamental frequencies of the order of 60 cycles or less, while many musical instruments can produce notes that have harmonics extending up to 15,000 cycles. Musical sounds are often accompanied by a noise caused by key clicks, hissing of air, etc. This noise represents energy more or less uniformly distributed over a range of frequencies at the high-frequency end of the audible spectrum. The results of an extensive investigation of the audible frequency range of musical sounds produced by typical instruments are shown in Fig. 445.[1] These were obtained from listening tests in which the frequencies at one end of the spectrum were progressively cut out until trained observers were able to detect a difference from the original sound. The results therefore represent the useful frequency range but do not necessarily mean that there is absolutely no energy outside the ranges indicated.

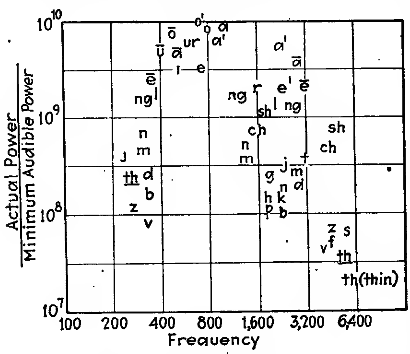


FIG. 444.—Chart showing intensities and region of most important frequency components of fundamental speech sounds. (*After Fletcher.*) When a sound has several principal components the intensity and frequency of each are indicated.

The power involved in musical sounds is often very large and in the case of large orchestras may reach peak values approaching 100 watts. The peak power depends, of course, upon the number and character of the instruments involved and upon the nature of the musical composition being rendered. Typical values of peak sound power obtained from musical instruments when played loudly are given in Table XV.[2]

[1] See W. B. Snow, Audible Frequency Ranges of Music, Speech, and Noise, *Bell System Tech. Jour.*, vol. 10, p. 616, October, 1931.

[2] See L. J. Sivian, H. K. Dunn, and S. D. White, Absolute Amplitudes and Spectra of Certain Musical Instruments and Orchestras, *Jour. Acous. Soc. Amer.*, vol. 2, p. 330, January, 1931.

The instruments producing the largest power are the bass drums, which generate peaks reaching 25 watts (approximately 3,000 times the peak power of the average voice), the pipe organ, the snare drum, the cymbals, and the trombone.

The way in which the sound power produced by musical instruments is distributed over the frequency spectrum depends upon the instrument involved and upon the fundamental frequency being played. Most, although not all, musical instruments are so constructed as to produce more sound power on the lower notes (*i.e.*, those below 500 to 1000 cycles)

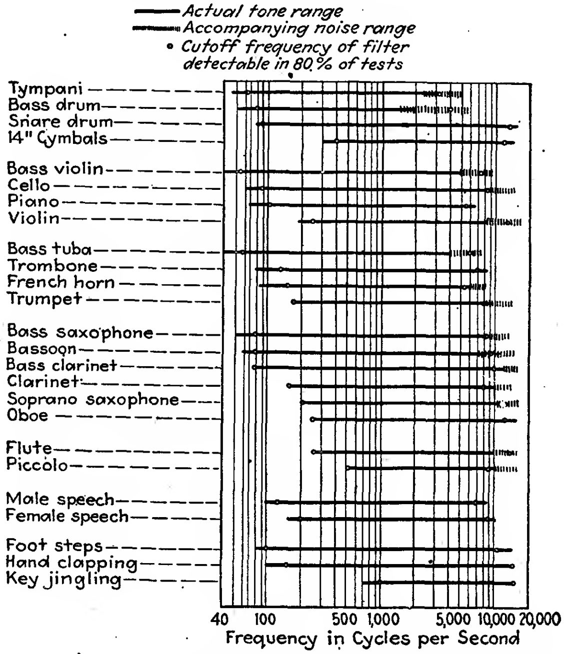


FIG. 445.—Frequency range of representative musical instruments as determined by listening tests (as given by W. B. Snow).

than on the higher frequencies, as is evident in Table XV. This table also gives the percentage of time intervals of ⅛-sec. duration in which the peak power is at least one-fourth of the maximum power observed, and it is apparent that different instruments vary greatly in this respect. The range of sound power encountered in the rendition of musical pieces depends upon the selection involved and upon the number of instruments participating. In the case of a symphony orchestra the sound power during the loudest passages may reach peaks that are 10,000,000 times as great as the sound power during the softest passages.

Noise.—Noise represents sounds in which the energy is more or less uniformly distributed over a considerable frequency range without a

definite pitch being present. In a crude way noise can be considered as consisting of a mixture of sound of all frequencies. Noises differ in the way in which their energy is distributed with frequency, as is apparent from Fig. 446. Many noises, such as the jingling of keys,

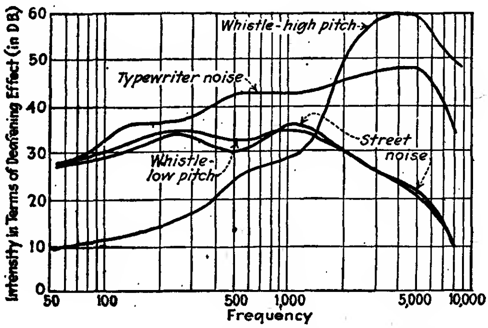


FIG. 446.—Energy distribution of typical noises expressed in terms of the deafening (or masking) effect. The distinguishing feature of noises is the more or less continuous energy distribution over a wide frequency range.

clapping of hands, and footsteps, contain important components having frequencies above 8500 to 10,000 cycles and do not sound natural when these are suppressed (see Fig. 445).

TABLE XV

Instrument	Peak power, watts	Percentage of ⅛-sec. intervals in which power is at least one-fourth of peak	Band containing maximum peaks, C.P.S.
36- by 15-in. bass drum	24.6	6	250 to 500
Snare drum	11.9	2½	250 to 500
15-in. cymbals	9.5	7½	8,000 to 11,300
Bass saxophone	0.288	25	250 to 500
French horn	0.053	6	250 to 500
Piccolo	0.084 0.021	½ 10	2,000 to 2,800
Piano	0.267	16	250 to 500
75-piece orchestra	13.4 66.5	9 1	250 to 500 8,000 to ∞

148. Characteristics of the Human Ear.[1]—The properties of the ear are of fundamental importance in sound work since it is through the medium of the ear that sound waves are observed. The frequency and

[1] Much of the material in this section is summarized from Fletcher, *loc. cit.* For further information also see Sec. 9 of "Electrical Engineers Handbook—Communication," John Wiley & Sons, New York.

amplitude ranges over which a normal ear receives auditory sensations are illustrated in Fig. 447. Frequencies below about 20 cycles are perceived by feeling rather than hearing, while frequencies above about 20,000 cycles are not heard by most ears. The limit marked "Threshold of feeling" in Fig. 447 represents the point at which the sound intensity becomes great enough to produce a sensation of pain, while the limit marked "Threshold of audibility" represents the minimum sound audible to the normal ear. It is apparent that the sensitivity of the ear depends upon the frequency and is maximum in the range 1000 to 3000 cycles.

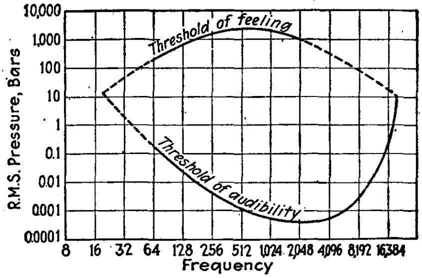


FIG. 447.—Average auditory sensation area of normal ears.

The smallest variation in sound amplitude that the ear is able to perceive is roughly a constant percentage of the original intensity, which is equivalent to stating that the loudness of a sound as measured by the ear is proportional to the logarithm of the sound intensity. The minimum percentage of change in sound energy that is detectable is in the order of 25 per cent (1 db) in the middle range of frequencies at moderate intensities, and becomes less when the intensity is low or when the frequency is high or low.

The wide range of intensities to which the ear responds and the fact that the sensitivity of the ear varies logarithmically make it convenient to use a logarithmic scale for measuring sound intensities. The unit commonly employed for this purpose is the decibel (abbreviated db), which is described in Sec. 54, and is ten times the common logarithm of a power ratio.

The minimum perceptible change in frequency varies with pitch, but over the frequency range 500 to 8000 cycles is very close to 0.03 per cent, with larger values for very low frequencies and for low sound levels.

The ear has a non-linear response to sound waves of large amplitude. The result is that with powerful sound waves the ear produces harmonics, as well as sum and difference tones, which are not present in the original sound and yet which are actually present in the hearing organs and are perceived by the brain. Such frequencies produced in the ear are called *subjective tones* and explain a number of sound phenomena. Thus the pitch of a sound is not changed by removing the fundamental frequency since the harmonics combine in the ear to produce a difference frequency that recreates the fundamental component in the form of a subjective tone. This non-linear character of the ear also makes many radio receiv-

ers and loud-speakers at least passably acceptable by regenerating in the form of subjective tones the low frequencies that the equipment itself fails to reproduce.

Another important consequence of the non-linear character of the ear is the phenomenon known as *masking*, which appears as a deafening to high-frequency sounds caused by the presence of a lower pitched sound. Masking arises from the fact that, when the ear produces harmonics of the low frequencies, these harmonics interfere with the perception of the higher pitched sounds, which are then said to be masked. Masking is particularly important in noisy locations, since it is equivalent to a deafening. It is the reason that it is necessary to raise the voice when carrying on a conversation in a noisy location.

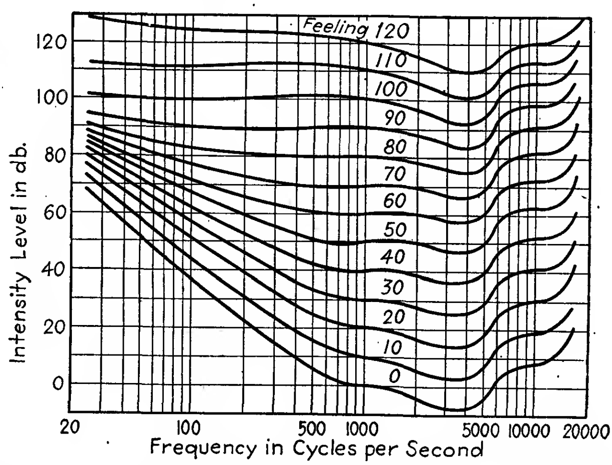


Fig. 448.—Contours giving intensity required for equal loudness as a function of frequency for pure tones. The numbers on the contours give the strength of an equally loud 1000-cycle tone in decibels above the minimum audible 1000-cycle tone.

Loudness.[1]—The magnitude of the auditory sensation produced by a sound is termed its *loudness*. Experiment shows that the relative loudness of different frequencies depends upon the absolute as well as upon the relative intensity. This is shown by Fig. 448, which presents experimental curves giving intensity level of the sound required for equal loudness as a function of frequency for the case of pure tones. It is seen from these curves that, when the intensity level is lowered, as, for example, when a radio is played softly, the lower frequencies tend to disappear. Thus, if 100- and 1000-cycle tones both have intensities corresponding to the 70-db contour in Fig. 448, reducing the intensity

[1] For further information see Fletcher and Munson, Loudness, Its Definition, Measurement, and Calculation, *Jour. Acous. Soc. Amer.*, p. 82, October, 1933.

of both by 40 db will reduce the loudness of the 1000-cycle tone to the 30-db contour, while the 100-cycle tone will be just at the limit of audibility and so will have disappeared.

149. Elements of Acoustics.[1]—The sound reaching an observer will generally differ from the sound as generated because of reflections from near-by objects. Consider, for example, the situation illustrated in Fig. 449, which shows only a few of the paths by which sound produced in a room may travel from source to observer. The direct route involving no reflections is the most important individual path, but, unless the observer is very close to the source, or unless the walls are lined with sound-absorbing material, large amounts of sound energy will reach the observer by way of the longer indirect paths involving reflections from the bounding surfaces.

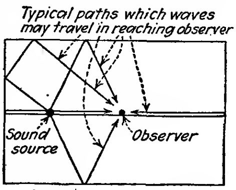


FIG. 449.—Diagram illustrating a few of the many routes which sound produced in a room may travel in reaching a listener.

The principal effects that these reflections have on the sound as observed are as follows:

1. The average intensity of the observed sound is raised because sound originally sent out in other directions is reflected back to the observer.
2. The relative amplitudes of the different frequency components of the sound may be altered as a result of selective absorption of the reflecting surfaces, which usually tend to reflect low frequencies more efficiently than high frequencies.
3. The relative amplitudes of the different frequency components of the sound will always be altered as a result of interference effects resulting from the fact that the phase with which the energy traveling along the different possible paths combines depends upon the position of the observer and upon the frequency.
4. The observed sound persists for some time after the original sound has ceased as a result of the greater time it takes the sound traveling along the indirect routes to reach the observer. This effect is known as reverberation.

The magnitudes of the first three of these effects depend primarily upon that fraction of the total energy reaching the observing point which has traveled an indirect path, and this in turn is determined by the relative lengths of the direct and the indirect paths and the fraction of the sound-wave energy that is absorbed upon reflection. When the direct path is short, or when the bounding surfaces are of such a character as to absorb a large fraction of the energy of sound waves striking them, most of the energy reaching the observer travels along the direct path, and interference, selective absorption, etc., are not important.

[1] For further information on the subject of acoustics, the reader should consult Sec. 9 "Electrical Engineers Handbook—Communication," John Wiley & Sons, New York.

When the observer is at a reasonable distance from the sound source, interference effects will produce considerable quality distortion. The extent of the quality distortion that can be expected is indicated in Fig. 450, which shows sound powers actually observed by a microphone in a small-sized classroom.[1] It will be noted that the relative sound intensity varies widely with frequency and microphone position. The result is that the actual sounds observed depend to a large extent upon the position of the observer.

Reverberation.—Reverberation in a room depends upon the ratio of the enclosed volume to the area of the bounding surfaces, and upon the average coefficient of sound absorption of the walls. It is measured in terms of the time after the sound source has been silenced that it takes a sound uniformly distributed throughout the room and having 1,000,000 times the minimum audible energy to die down to inaudibility. This reverberation time depends upon the volume of the space involved,

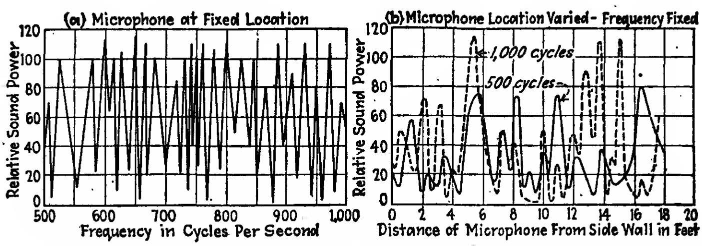


FIG. 450.—Variations in observed sound power that result from changes in position and changes in frequency because of interference effects between standing wave trains existing in small classroom.

the area of sound-absorbing surface present (including walls, furniture, people, etc.), and the average coefficient of sound absorption of the surfaces. The reverberation time is commonly of the order of several seconds in large theaters and auditoriums where the sound travels long distances between reflections. In living rooms of ordinary homes the fact that the sound waves have opportunity to travel only a short distance between reflections results in rapid absorption and hence a very short reverberation time.

It might be thought that the smaller the reverberation time the better, but this is not necessarily true because the ear normally expects a certain amount of reverberation and because reverberation enhances certain musical and oratorical effects. Thus with music the presence of reverberation helps the different players in an orchestra to play together properly and increases the effectiveness of passages intended to convey

[1] The writer is indebted to P. G. Caldwell, formerly graduate student in electrical engineering at Stanford University, for these curves.

the impression of power. Similarly in speech the presence of reverberation increases the average intensity of the sound above the level that would be present outdoors or in a room with perfectly absorbing walls, and thereby makes it possible for a speaker's voice to reach a larger number of people.

The optimum reverberation time for a broadcast studio varies with different conditions, but is always less than for the corresponding theater or auditorium. This is because the ultimate listener receives reverberation from both the broadcast studio and the room in which the sound is reproduced.

Experience indicates that the optimum reverberation time for broadcast work is about two-thirds of that considered best for ordinary speech or music rooms. This result can be most satisfactorily achieved either by using a directional microphone so placed as to discriminate against sound reflected from room boundaries in favor of sound coming directly from the source, or by using a relatively long narrow studio having high absorption at one end (where the microphone is located) and low absorption at the other (where the sound is produced). These expedients reduce the amount of reverberation picked up by the microphone, while allowing the performers to work in the presence of normal reverberation.

The reverberation time can be controlled by the use of sound absorbants, and studios, auditoriums, etc., must be acoustically treated so that the optimum reverberation time is obtained. This is done by use of acoustic tiles and plasters on walls and ceiling to provide a certain minimum of absorption, and by adding rugs, drapes, etc., when circumstances call for still less reverberation.

150. Effects of Distortion in the Reproduction of Sound.—In designing a system for the electrical reproduction of sound it is necessary to consider the various ways in which the original sound may be distorted, and the consequences of this distortion. The most important single factor involved is the frequency range of the electrical system. The frequency band required for substantially perfect reproduction varies with the nature of the sound involved, as is apparent from Fig. 445, but is about 100 to 10,000 cycles for speech and 60 to 15,000 cycles for music. Experience indicates, however, that a frequency range of 80 to 8000 cycles is adequate to give excellent reproduction of all sounds except certain types of noises. The usual broadcast program has a frequency range of the order of 100 to 5000 cycles, and understandable although not natural speech can be transmitted with a much narrower range of frequencies, as illustrated by the telephone, which has a frequency range of 250 to 3000 cycles.

Experimental investigations of the effect of amplitude distortion, such as introduced by an overloaded amplifier, indicate that, when the

full frequency range is reproduced, distortions of 3 to 5 per cent are detectable, and that 10 per cent is very noticeable.[1] Odd harmonics are more troublesome than even harmonics, and the amount of distortion permissible becomes less as a wider frequency range is reproduced.

The intensity level at which the sound is reproduced is also of considerable importance. If the reproduced sound is at lower intensity than the original, the low frequencies appear to be weaker than they should, as explained in connection with Fig. 448, while the tendency of the high frequencies to be masked by the lower frequencies is reduced. Both of these effects work in the same direction, and combine to make it necessary to vary the frequency response with volume level if high-quality reproduction is to be obtained. The "tone-compensated" volume control used in many radio receivers is for the purpose of providing this type of correction.

The acoustics of the space where the microphone is located and of the room containing the loud-speaker affect the fidelity of the reproduced sound as explained above. In this connection it is to be noted that a person listening to the original sound is able, as a result of having two ears, to discriminate against noise and reverberation arriving from directions other than along the direct path, whereas the microphone cannot. The result is that studio reverberation and noise picked up by the microphone and reproduced by a loud-speaker do not sound the same as normal reverberation and noise. Ordinary reproducing systems also fail to give space relationships, since all the reproduced sound comes to the observer from the same direction, irrespective of the location of the original sound source with respect to the microphone.[2]

The extremely wide range of intensities encountered in ordinary sound makes it necessary in electrical reproducing systems to reduce the intensity of the very loudest passages to prevent overloading amplifiers, etc., and to increase the intensity of the weaker passages to prevent their being lost in the background noise. This introduces a form of distortion to which the musically inclined tend to be particularly sensitive, but which can be readily remedied by means of automatic volume expansion, as described in Sec. 48.

The exact phase relations between the various components of a complex sound are unimportant unless the phase shift is great enough to produce an appreciable time-delay distortion, such as 0.01 sec. Varia-

[1] See Frank Massa, Permissible Amplitude Distortion of Speech in an Audio Reproducing System, *Proc. I.R.E.*, vol. 21, p. 682, May, 1933.

The percentages given are the percentages of harmonics that result with a sine wave having an amplitude equal to the peak amplitude of the complex audio wave.

[2] The illusion of space can be obtained by the use of two or three separate channels. See a series of six papers on Auditory Perspective, *Electrical Eng.*, January, 1934.

tion in the phase (such as $\pm 180°$) of various components with respect to each other is not even detectable by the ear.[1]

In view of all these ways in which the original sound may be distorted, it might be thought rather remarkable for the reproduced sound to be acceptable. Actually, however, the ear is not a particularly critical device, and is reasonably content with rather imperfect results. This is in part due to the fact that the ear has been trained since infancy to accept many forms of distortion; for example, the distortion discussed in connection with Fig. 450 is accepted as perfectly natural.

Articulation.—The extent to which a transmission system is capable of reproducing the original speech meaning is measured in terms of the *articulation.*[2] Articulation is ordinarily based upon the accuracy with which the fundamental voice sounds are perceived, and can be tested using random combinations of vowel and consonant sounds in vowel-consonant, consonant-vowel, and consonant-vowel-consonant combinations. An articulation of 90 per cent on such a test would mean that 90 per cent of the individual voice sounds were correctly received.

The results of a typical articulation test are shown in Fig. 451. It will be noted from this figure that, while most of the energy of speech is in the lower frequency range, the higher frequencies are essential to satisfactory articulation. Thus, when all frequencies above 1550 cycles are suppressed, the articulation is reduced to 65 per cent although less than 12 per cent of the sound energy is eliminated, while suppressing all frequencies below 1550 cycles gives an articulation of 65 per cent even though only about 12 per cent of the original sound energy remains. In this connection it must be kept in mind that naturalness and articulation are not synonymous. Thus Fig. 451 shows that the articulation is not appreciably reduced by suppressing all frequencies below 500 cycles, although doing so destroys the naturalness of the voice. In general, satisfactory naturalness requires the preservation of a much greater frequency range than is needed from the standpoint of mere articulation.

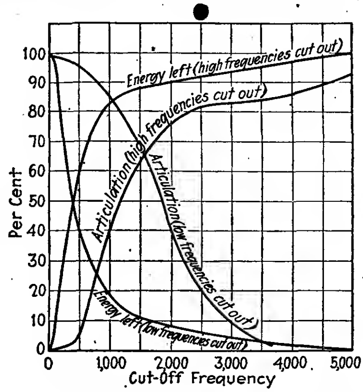


FIG. 451.—Variation of articulation for voice sounds as one end of the frequency spectrum is cut out, together with the fraction of the original sound power remaining after the restriction of the frequency band.

[1] See Balth. van der Pol, A New Transformation in Alternating-current Theory with an Application to the Theory of Audition, *Proc. I.R.E.*, vol. 18, p. 221, February, 1930.

[2] See Fletcher, *loc. cit.*, for more detailed information on articulation and articulation testing.

151. Dynamic Loud-speakers Employing Paper Cones.—Nearly all loud-speakers employed in radio receivers are of the "dynamic" type illustrated in Fig. 452. This consists of a paper cone to the apex of which is fastened a coil (commonly called the *voice coil*) located in a strong magnetic field and carrying the audio-frequency currents to be transformed into sound waves. In such an arrangement the action of the magnetic field on the coil current produces a mechanical force that vibrates the paper cone and causes the radiation of sound waves. The cone is supported in some manner around its outer edge, while the coil is held in position and supported by means of a flexible spider. The entire coil and cone assembly is therefore free to move as a unit, and under ordinary conditions is proportioned in such a manner as to have a resonant frequency at the lower end of the frequency range to be reproduced. The cone must be mounted in a baffle as shown in Fig. 452, or in a box, in order to prevent the radiation from the front and back sides from canceling at low frequencies.

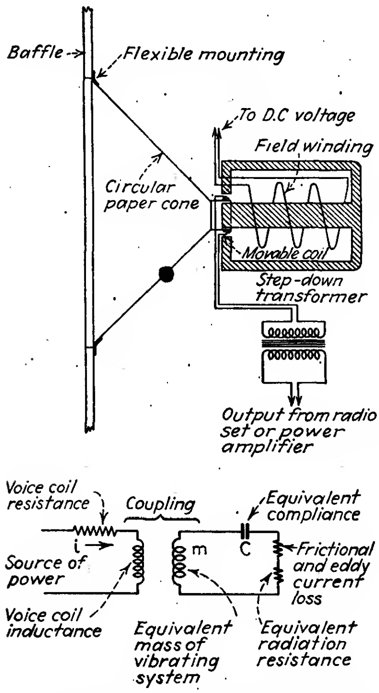


FIG. 452.—Cross section of typical dynamic type of loud-speaker, together with equivalent electrical circuit of electromechanical system.

The characteristics of a typical high-grade dynamic speaker are shown in Fig. 453. The frequency range is adequate for ordinary broadcast receivers, and the response is sufficiently uniform over this frequency range to sound entirely satisfactory upon listening tests.

Analysis of Dynamic Speaker Action.[1]—The force that is exerted upon the voice coil as a result of the action of the magnetic field upon the voice-coil currents acts against a mechanical impedance consisting of a mass, a compliance, and a resistance. The mass consists of the effective mass that the coil and cone assembly offers to the frequency involved plus the fluid mass caused by the air in contact with the cone. The equivalent compliance is determined primarily by the spider supporting the voice

[1] For further information see H. Olson and F. Massa, "Applied Acoustics," P. Blakiston's Son & Company, Philadelphia; Frank Massa, Loud Speaker Design, *Electronics*, vol. 9, p. 20, February, 1936; J. D. Seabert, Electrodynamic Speaker Design Considerations, *Proc. I.R.E.*, vol. 22, p. 738, June, 1934.

coil, but is also influenced by closed air spaces, etc. The effective resistance to motion includes eddy-current losses, sound energy that is radiated, etc. The velocity of the resulting voice-coil vibration is proportional to the force divided by the mechanical impedance, while the amount of sound energy radiated is proportional to the square of the velocity times the radiation component of the resistance.

The voice coil in vibrating cuts across the direct-current magnetic field of the loud-speaker and so has a back voltage induced in it. This causes a voltage drop that is equivalent to adding an impedance to the voice-coil circuit, and, since this additional impedance is caused by vibration, it is termed *motional impedance.* The real and reactive energies represented by the voice-coil current flowing through the voice-

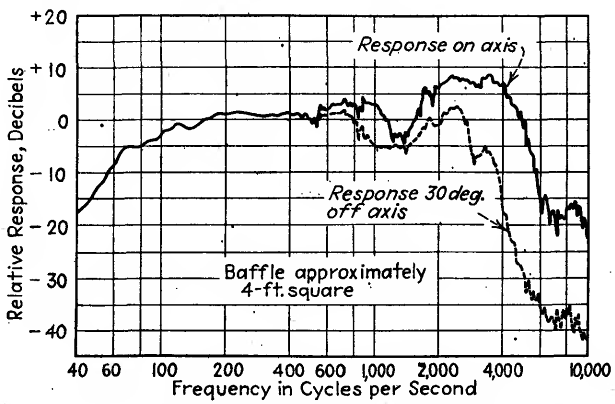


FIG. 453.—Response curve of typical cone dynamic speaker.

coil motional impedance are the real and reactive energies, respectively, which the electrical circuits deliver to the vibrating system in order to sustain the vibrations.

The quantitative relations existing in the electromechanical system represented by the dynamic loud-speaker can be summarized by the following equations. The force that the magnetic flux of the field circuit exerts on the voice coil is

$$\text{Force in dynes} = \frac{Bli}{10} \tag{253}$$

The effective impedance against which this force is exerted is

$$\left.\begin{array}{l}\text{Impedance of cone and voice coil in}\\ \text{mechanical ohms or dyne sec/cm}\end{array}\right\} = Z = R + j\left(\omega m - \frac{1}{\omega C}\right) \tag{254}$$

The velocity with which the voice coil moves is equal to the force divided by the impedance, or

$$\text{Velocity in cm per sec.} = \mu = \frac{\text{force in dynes}}{\text{impedance in mechanical ohms}} \tag{255}$$

The amplitude of the vibration is

$$\text{Amplitude of vibrations in cm} = \frac{\text{velocity in cm per sec}}{\omega} \tag{256}$$

The energy radiated per square centimeter of cone surface is

$$\text{Radiated energy in watts per sq cm} = \mu^2 R_a \times 10^{-7} \tag{257}$$

The motional impedance is the counter e.m.f. E_c induced in the voice coil divided by the voice-coil current causing the vibration, or

$$\left.\begin{matrix}\text{Motional impedance}\\ \text{in ohms}\end{matrix}\right\} = Z_m = \frac{E_c}{i} = \frac{B^2 l^2}{Z} \times 10^{-9} \tag{258}$$

The notation in these equations not given in connection with the equations is

B = air-gap flux density in gausses
l = length of wire in the voice coil in centimeters
i = voice-coil current in amperes
R = total resistance of cone in mechanical ohms (dyne seconds per centimeter) including radiation loss, eddy-current loss, frictional loss resulting from the bending of the cone material, etc.
R_a = radiation resistance per square centimeter of cone area
m = effective mass in grams which vibrating system offers driving force, including coil mass, effective cone mass, and effective mass of air load
C = compliance of vibrating system in centimeters per dyne
ω = 2π times frequency.

A study of Eqs. (253) to (258) shows that, as far as the electrical circuits are concerned, the mechanically vibrating system is equivalent to a tuned secondary. This leads to the equivalent electrical circuit for the dynamic loud-speaker shown in Fig. 452, in which the mechanical force acting on the voice coil is represented by the voltage induced in the secondary, while the motional impedance is the coupled impedance. It will be noted that the mechanical impedance of the secondary is of the same form as that of a tuned circuit, and that the velocity of vibration is equivalent to the current in the electrical analogue. The energy dissipated in the secondary resistance of the equivalent electrical circuit hence represents the energy required to maintain the cone vibrations, and the fraction of this energy dissipated in the portion of the secondary impedance corresponding to radiation resistance represents the amount of energy actually converted into sound power.

At low frequencies (commonly up to 500 to 1000 cycles) the paper cone acts approximately as a piston diaphragm having a diameter equal to the diameter of the cone.[1] Under these conditions the presence of the air in contact with the vibrating diaphragm produces a mechanical radiation resistance R_a per unit area which varies with frequency as shown in Fig. 454. For frequencies low enough so that the diameter of the cone is appreciably less than a half wave length it is seen that the radiation resistance is indirectly proportional to the square of the frequency, so the total sound power will be independent of frequency if the velocity is inversely proportional to frequency. This result can be readily realized by making the resonant frequency of the cone and coil assembly

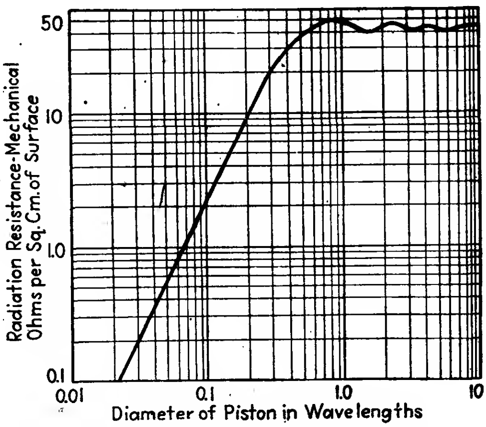


FIG. 454.—Radiation resistance of piston radiator per square centimeter of surface. (When both sides of piston radiate, the total surface is twice area of one side.)

less than the lowest frequency to be reproduced, since then the principal impedance to motion in Eq. (254) is supplied by the inertia of the coil mass and is proportional to frequency.

At the resonant frequency of the coil and cone combination, the impedance to motion becomes very small, so that there is a tendency for the velocity of vibration and hence the radiated power to be large. However, under these conditions the motional impedance is also large, and examination of the equivalent electrical circuit of Fig. 452 shows that this exerts a counteracting tendency by reducing the current through

[1] The assumption of piston action is actually only a rough approximation, since studies show that the actual cone motion is of a very complex character even at very low frequencies. The observed behavior is reasonably consistent with piston action at low frequencies, but it is probable that departures from true piston action begin to be appreciable at a few hundred cycles and become progressively more pronounced as the frequency is increased, until at about 1000 cycles the vibrations are in the form of waves radiating out from the apex of the cone.

the voice coil. By proper design, particularly by use of a magnetic field of appropriate strength, it is possible to make the response at the resonant frequency substantially the same as at higher frequencies, or to have either a peak or a dropping off.[1] A strong resonant peak is sometimes deliberately introduced to increase the sound power radiated at very low frequencies and thus to give the appearance of a strong bass. This practice is not very desirable, however, since pronounced transient oscillations at the resonant frequency will then take place upon shock excitation. Below resonance the response drops off rapidly as a result of the rapid rise in elastic impedance offered to the driving force.

At frequencies so high that the diameter of the cone exceeds a half wave length the radiation resistance becomes substantially independent of frequency. With piston action and a velocity inversely proportional to frequency, the radiated power is then inversely proportional to the square of the frequency. This tendency for the response to fall off at high frequencies can be counteracted by so designing the paper cone that it ceases to operate as a piston at these higher frequencies. The cone vibrations are then in the form of waves traveling outward from the apex, with the result that the center part of the cone vibrates much more intensely than the outer edges. As the frequency is increased, the action is therefore very much as though the size of the cone were progressively decreased. This reduces the effective mass, increasing the velocity above what it would be with piston action and increasing the total radiated sound energy. The power still tends to drop somewhat at the higher frequencies, but, since there is a tendency for the radiated energy to be concentrated more nearly along the axis of the cone the higher the frequency, it is found that the response directly in front of the cone can be made substantially constant up to rather high frequencies, although there is a falling off of the higher frequencies to the side (see Fig. 453).

[1] This assumes that the internal impedance of the power source indicated in Fig. 452 is equal to or less than the equivalent load impedance offered by the voice coil, as is the case when a triode power tube is used to drive the speaker. With pentode tubes, where the plate resistance is much higher than the load impedance, the high back voltage at resonance merely increases the voltage drop across the speaker input without changing the current. The result is then an accentuation of the resonance effect since this increase in power input comes when there is already a tendency for the output to be excessive. This fact that the pentode does not help damp out speaker resonance as does a triode tube has until recently been considered a major defect of the pentode power tube. The development of the feed-back amplifier has altered the situation, however, since feedback circuits such as that of Fig. 143*c* which act to stabilize the output voltage operate to damp out resonances just as though the tube had a low plate resistance. It is to be noted, however, that other feedback circuits such as Fig. 143*a* which act to stabilize the output current have just the opposite effect, tending to give the effect of an excessively high plate resistance even when the tube itself has a low plate resistance.

The ultimate limit to the high-frequency response is set by the mass of the voice coil.

Miscellaneous Considerations.—A study of the factors involved in the behavior of a cone speaker shows that a large cone and a proportionately large voice coil are best for the radiation of low frequencies, but that a small cone and correspondingly light voice coil are favored for the reproduction of the higher frequencies. As a result, a very wide frequency range is best reproduced by a pair of loud-speakers, one with a large cone for the lower frequencies and the other with a small cone and light voice coil for the higher frequencies, with accompanying electrical networks for separating the high- and low-frequency audio currents.[1] An ingenious combination of a low-frequency speaker and a high-frequency speaker in one unit is provided by the articulated voice-coil arrangement of Fig. 455.[2] Here the speaker is provided with two voice coils, one heavy and one light, which are connected together by an elastic link as shown. At low frequencies both coils carry current and act together as a unit, but at high frequencies the large mass of the heavy voice coil causes it to remain substantially motionless while the light voice coil continues to function as a result of the elastic compliance C_{M_1} provided between the coils. The low-frequency coil is also shunted by a condenser in order to by-pass the impedance of the large coil to high frequencies.

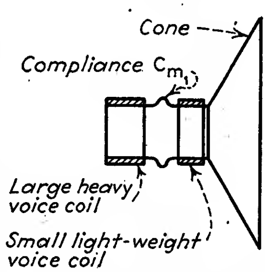


FIG. 455.—Articulated voice coil arrangement for extending high-frequency response.

The efficiency of a loud-speaker in transforming electrical energy into acoustical radiation is relatively low except at the resonant frequency because the coupling between the air and the vibrating cone is so poor that very little of the driving force is actually used in doing work against the air. Under practical conditions the average efficiency of a cone speaker is less than 5 per cent.

The paper cone must be mounted in some form of baffle as shown in Fig. 452 in order that the waves radiated from the two sides of the cone, which are produced with a phase difference of 180°, will not cancel each other. The baffle diameter should be of the order of a half wave length at the lowest frequency for which little or no loss in the sound output is desired, and for best results should have an irregular outline in order to eliminate the possibility of destructive interference between front and

[1] The design of such networks is described by John K. Hilliard and Harry R. Kimball, Dividing Networks for Loud Speaker Systems, *Technical Bulletin, Academy of Motion Picture Arts and Sciences*, Mar. 3, 1936.

[2] See Harry F. Olson, A New Cone Loud Speaker for High Fidelity Sound Reproduction, *Proc. I.R.E.*, vol. 22, p. 33, January, 1934.

back side radiation at certain critical frequencies. In the usual radio receiver the cabinet supplies the baffle, and where good response down to low frequencies is desired the cabinet must be of ample size, or must be provided with some form of acoustic labyrinth or resonator as described in Sec. 153.

152. Horns.[1]—The horn is essentially an acoustic coupling device which transforms acoustic energy at a high pressure and low velocity to energy at a low pressure and high velocity. A horn can therefore be used to increase the load that the air produces upon the driving mechanism of a loud-speaker, and thereby makes it possible to obtain a better match between a relatively heavy driving mechanism and the fluid air.

A typical horn speaker is illustrated schematically in Fig. 456, while the details of a representative moving-coil driving arrangement are

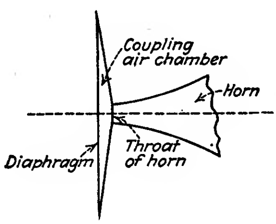


FIG. 456.—Diaphragm, throat, and coupling air chamber of horn-type loud-speaker.

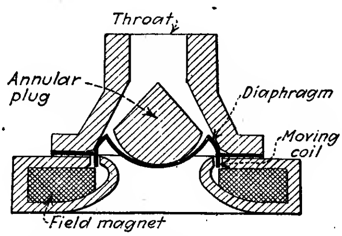


FIG. 457.—Moving-coil type of driving unit used in a commercial horn-type loud-speaker.

shown in Fig. 457. Sound waves produced at the throat by the diaphragm vibrations travel along the horn, expanding in an orderly manner until large enough to transfer their energy to space without undue disturbance.

Design of Horns.—The factors involved in the design of a horn are the taper, the mouth area, and the throat area. For proper operation the taper should be such that the cross-sectional area is proportional to an exponent of the distance along the horn. That is

$$\text{Area at distance } x \text{ from throat} = A_0\epsilon^{Bx} \tag{259}$$

where A_0 is the throat area and B is a constant that determines the rate

[1] For more detailed information on horns and horn-type loud-speakers, see C. R. Hanna and J. Slepian, The Function and Design of Horns for Loud-speakers, *Trans. A.I.E.E.*, vol. 43, p. 393, 1924; C. R. Hanna, Loud-speakers of High Efficiency and Load Capacity, *Trans. A.I.E.E.*, vol. 47, p. 607, April, 1928; E. C. Wente and A. L. Thuras, A High-efficiency Receiver for a Horn-type Loud-speaker of Large Power Capacity, *Bell System Tech. Jour.*, vol. 7, p. 140, January, 1928; Olson and Massa, *loc. cit.*; Massa, *loc. cit.*

at which the horn opens out.[1] An infinite horn when tapered in this way transmits freely all frequencies appreciably above a certain critical "cut-off" frequency, and nothing below. This transmission characteristic depends upon the ratio $2\pi f/B$ as shown in Fig. 458. Cut-off occurs when $f = 2730B$, and the transmission falls off rapidly for frequencies less than $4000B$, where B is evaluated from Eq. (259) when dimensions are in centimeters. The proper taper for an exponential horn is therefore determined by the lowest frequency sound that is to be radiated.

The mouth area of a horn determines the lowest frequency sound wave that can be transferred from the horn to free space without setting up resonances in the air column. When the mouth diameter is $2\lambda/3$, such resonances are negligible, and do not become excessive until the diameter is less than $\lambda/4$, where λ is the wave length of the sound. The mouth of a horn is hence determined by the lowest frequency that is to be handled.

The throat area of the horn determines the loading that the horn places upon the diaphragm of the driving mechanism. When the diaphragm has the same area as the throat, the acoustic resistance that is offered to the vibrations of the diaphragm is 41.5 mechanical ohms per square centimeter, while, if the throat represents a constriction, the acoustic resistance offered to the diaphragm is $41.5(S_d/S_0)^2$ mechanical ohms per square cm. of throat, where S_d is the diaphragm area, and S_0 is the area of the throat. A small throat therefore increases the acoustic loading on the diaphragm, and so raises the efficiency with which the electrical energy is transformed into sound energy. However, a small throat has various disadvantages, such as requiring a longer horn, increasing the frictional losses, distortion due to excessive throat pressures, etc., so that in actual practice the throat size is chosen as a compromise between various factors.

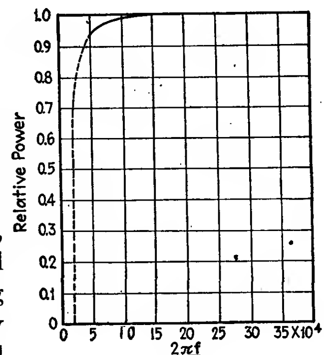


FIG. 458.—Relative power transmitted along an exponential horn with different values of taper (all lengths measured in centimeters).

It is to be noted that the length of a horn is determined by the mouth, the throat, and the taper.

Performance of Horn Speakers.—The low-frequency limit of a horn speaker is set by the taper and mouth of the horn, as has been explained. The ultimate high-frequency limit is set by the mass of the moving coil

[1] Other forms of taper, such as conical, are occasionally employed in horns but have the disadvantage of much less uniform transmission with variation of frequency than the exponential horn.

and diaphragm. In order to realize this possible high-frequency limit, it is, however, necessary that the distance from the throat to various parts of the diaphragm vary by less than a half wave length, since otherwise the sound waves produced by various parts of the diaphragm will not add up in phase at the throat and there will be a falling off in response at the higher frequencies. In order to minimize difficulties of this sort, it is common practice to employ an annular plug between diaphragm and throat to equalize the distance to the throat from various parts of the diaphragm as shown in Fig. 457.

The sound power generated by the diaphragm of a horn speaker is proportional to the square of the diaphragm velocity. In order to obtain a frequency response that is substantially independent of frequency, it is accordingly necessary that the impedance which the vibrating system offers to its driving force be substantially resistive. This condition is realized by making the acoustic loading on the diaphragm high and by designing the vibrating system to have the lowest possible mass together with a resonant frequency in the middle of the response range. Under these conditions the impedance that the diaphragm offers to its driving force is predominantly resistive, and the velocity is relatively independent of frequency.

The efficiency with which the electrical energy supplied to the voice coil is converted into acoustic energy depends upon the ratio of copper to radiation resistance, and becomes higher the greater the diaphragm loading. This indicates the desirability of a small throat; but at the same time the throat must not be so small that an appreciable part of the acoustic energy is lost in friction, or that the sound-wave pressure in the throat is sufficient to cause excessive distortion. Under practical conditions efficiencies of the order of 25 to 50 per cent are commonly obtained.

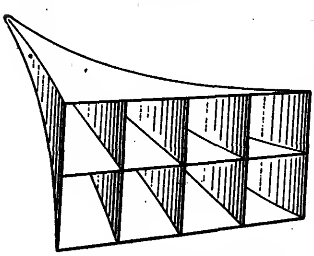


FIG. 459.—Horn with deflecting vanes to distribute high-frequency energy evenly.

A horn tends to concentrate the sound along its axis, with this tendency becoming more pronounced the higher the frequency. The result is that, when the total sound power generated is independent of frequency, the high frequencies are too strong along the axis and too weak to the side. This effect can be taken care of when necessary by providing the horn with deflecting vanes, as shown in Fig. 459, for the purpose of spreading the energy equally for all frequencies.

Directional Baffles.—In public-address and theater work it is common practice to employ ordinary cone speakers which are provided with a short horn, or *directional baffle,* as shown in Fig. 460. The distinguishing

feature of such an arrangement is a short horn, made possible by the use of a throat area equal to or only moderately less than the size of the paper cone, and the enclosed box for the purpose of preventing radiation from the back of the cone. At frequencies low enough so that the box dimensions are appreciably less than a quarter of a wave length, the presence of the box is equivalent to adding stiffness to the vibrating system. Resonances at higher frequencies are avoided by packing the box with felt or other acoustic absorbent and by taking advantage of the fact that the absorption coefficient increases with frequency.

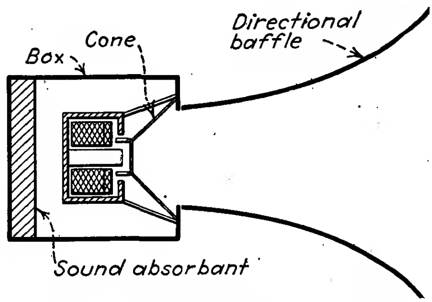


FIG. 460.—Cone speaker provided with directional baffle.

A directional baffle concentrates the sound and also increases the efficiency of radiation, particularly in the frequency range in which the cone functions as a piston. This is because with a throat of the same size or smaller than the cone the acoustic impedance will be at least

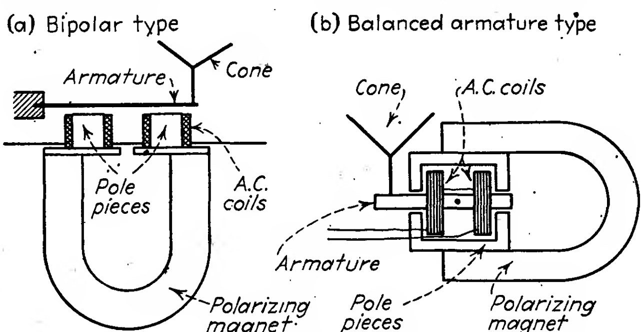


FIG. 461.—Typical magnetic-speaker driving systems.

41.5 mechanical ohms per square centimeter at low frequencies, whereas it is seen from Fig. 454 that the acoustic resistance is much less for the same cone operated in a plane baffle at a low frequency. In this way the acoustic loading can be increased to the point where it tends to control the impedance of the vibrating system at low and medium frequencies, with the result that efficiencies of the order of 25 per cent are typical.

153. Loud-speakers. Miscellaneous Considerations. *Magnetic Speakers.*—While most speakers employ a moving-coil driving mechanism, some of the very cheap loud-speakers employ driving systems such as illustrated in Fig. 461. These are termed magnetic speakers, and make use of what is essentially a large telephone-receiver type of

driving mechanism such as discussed in Sec. 154. The chief disadvantage of the magnetic speaker is that the force-displacement characteristic of the driving mechanism of magnetic speakers becomes non-linear when the amplitude is at all appreciable, so that the output obtainable without distortion is very low, particularly on the lower frequencies.

Distortion.—If driven too hard all loud-speakers overload and generate spurious frequencies. One form of distortion is caused by a non-uniform distribution of flux in the air gap in which the voice coil operates. Such distortion is most pronounced when the amplitude of vibration is large, and so is found at low frequencies, particularly the resonant frequency of the vibrating system. In improperly designed speakers in which excessive resonance is allowed in order to boost the low notes, it is not uncommon for the harmonics to reach 50 per cent to 100 per cent or even more of the fundamental.[1] Another important source of harmonics is a non-linear force-displacement characteristic of the moving system arising as the result of non-linear action in the paper cone, supporting spider, etc. Under certain conditions non-linear distortion in the cone also produces relatively strong sub-harmonics, *i.e.*, frequencies that are one-half, one-third, etc., of the driving frequency.

In horns, high air pressures produced in the throat cause distortion because the adiabatic compression of air is non-linear with respect to pressure. This effect tends to be most pronounced at the higher frequencies and sets a limit to the lowest permissible throat area in horns intended to handle high sound powers.[2]

Acoustic Labyrinths, Resonators, Etc.—The practical impossibility of obtaining sufficient baffle action from a receiver cabinet to reproduce the very low frequencies has led to the development of various expedients for meeting this problem. One such arrangement involves connecting the back of the loud-speaker to open space through a tortuous passage or labyrinth. This gives the effect of a large baffle at the very low audio frequencies while tending to absorb the higher frequencies. Such an acoustic labyrinth is shown in the large compartment below the receiver chassis in Fig. 320.

[1] Thus see Hugh S. Knowles, Loud Speaker Cost versus Quality, *Electronics*, vol. 6, p. 240, September, 1933.

[2] The resulting distortion can be approximated by the equation

$$\text{Percentage of second harmonic} = \frac{\sqrt{W}}{81}\left(\frac{f_1}{f_c}\right) \times 100 \qquad (260)$$

where W is the acoustic watts per square centimeter of throat area and (f_1/f_c) is the ratio of driving frequency to horn cut-off frequency. See A. L. Thuras, R. T. Jenkins, and H. T. O'Neil, Extraneous Frequencies Generated in Air Carrying Intense Sound Waves, *Bell System Tech. Jour.*, vol. 14, p. 159, January, 1935.

Another arrangement coming into use consists in inclosing the back of the cone by a box that is provided with a port that opens in the same direction as the front of the speaker. By proper proportions of the inclosed space, port size, and port position, it is possible so to shift the phase of the radiation from the back of the cone that the sound issuing from the port will be in phase with the radiation from the front of the cone at the lower frequencies. At higher frequencies this is not necessarily so, but by the proper use of absorbing material in the inclosed space, the high frequencies can be sufficiently damped to prevent excessive resonance.

Very low frequencies can also be reproduced by inclosing the back side of the cone speaker with a box which is normally lined with sound-absorbing material. This prevents radiation from the back of the cone, and so avoids the cancellation of the low frequencies when there is not sufficient baffle area. The effect of the enclosed space is to add stiffness to the equivalent vibrating system of the cone at low frequencies, while the higher frequencies are absorbed.

A still different method of attacking the problem of low-frequency reproduction consists in coupling the equivalent of a folded horn to the back side of a small cone by means of an air chamber of such size that at low frequencies the horn functions in a normal manner, while in the mid- and high-frequency range the horn ceases to function and the front side of the cone is depended upon for reproduction.[1]

High-frequency Speakers.—The efficiency, acoustic output, uniformity of response, etc., can all be improved by limiting the frequency range. This has led to the use of dual-speaker systems where high quality, or high power, or both, are important. The low-frequency unit of such an arrangement commonly consists of a horn speaker, or cone with directional baffle. The high-frequency unit always employs a horn, but the method of driving varies. The most widely used drives are a small paper cone with light voice coil, and a metal diaphragm and moving-coil arrangement similar to Fig. 457. Other arrangements that have been suggested include a piezo-electric driving arrangement.[2]

154. The Telephone Receiver.—The term *telephone receiver* is used here to denote those devices which convert electrical energy into sound waves and which are held against the ear when used. This is in contrast with loud-speakers, which are arranged to produce sound waves that spread over a large volume.

[1] See H. F. Olson and R. A. Hackley, Combination Horn and Direct Radiator Loud Speaker, *Proc. I.R.E.*, vol. 24, p. 1557, December, 1936.

[2] See Stuart Ballantine, A Piezoelectric Speaker for the Higher Audio Frequencies, *Proc. I.R.E.*, vol. 21, p. 1399, October, 1933.

All types of telephone receivers make use of a diaphragm that is effectively sealed to the ear by means of a vented cap, so that, as the diaphragm vibrates, the pressure of the small quantity of air trapped between the diaphragm and the ear drum varies in accordance with the displacement of the diaphragm. Investigations have shown that, in order to obtain distortionless reproduction, the amplitude of the diaphragm vibrations should be proportional to the current supplied to the telephone receiver and independent of the frequency of this current.

Magnetic-diaphragm Telephone Receiver.[1]—The type of telephone receiver most widely employed makes use of a permanent magnet, upon the pole tips of which are coils that carry the voice currents. The magnetic circuit is closed by means of a magnetic diaphragm, which is set in vibration by the audio-frequency currents passed through the receiver windings. The cross section of a typical telephone receiver of the type worn with a head band is shown in Fig. 462. The receivers used in the telephone system are similar except for the fact that the permanent magnet is located inside the hand piece.

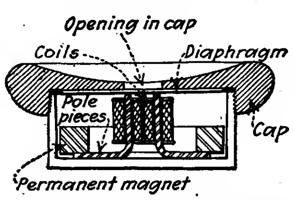


FIG. 462.—Cross section of typical watch-case type of telephone receiver.

The permanent magnet increases the sensitivity and prevents distortion. This results from the fact that the pull on each unit area of the diaphragm is proportional to the square of the flux density in the air gap. Thus, if B_o is the flux density produced by the permanent magnet and $B_s \sin \omega t$ is the flux density produced by the current in the receiver windings, then

$$\text{Pull on diaphragm} \propto (B_o + B_s \sin \omega t)^2 = B_o^2 + 2B_oB_s \sin \omega t + \frac{B_s^2}{2}(1 - \cos 2\omega t) \qquad (261)$$

The first term on the right-hand side of Eq. (261) represents the constant pull produced by the permanent magnet. The second term is a force that varies in accordance with the current passing through the receiver winding and is proportional to the strength of the permanent magnet. This is the force that produces the desired diaphragm vibrations. The final term consists of a constant pull and a double-frequency distortion alternating-current force, both of which are relatively small if the flux from the permanent magnet is large, but which in the absence of a permanent magnet represent the only forces exerted on the diaphragm.

[1] An exhaustive treatment of the magnetic-diaphragm telephone receiver is given by A. E. Kennelly, "Electrical Vibration Instruments," The Macmillan Company, New York.

The diaphragm of a telephone receiver has mass, elasticity, and friction, and so behaves under the influence of an applied force according to exactly the same laws as the moving system of a dynamic speaker [see Eqs. (253) to (256)]. Since the response is proportional to the amplitude of vibration, reference to Eqs. (255) and (256) shows that for a constant coil current the response is independent of frequency up to the region of diaphragm resonance, where the response depends upon the damping. At higher frequencies the response falls off rapidly. For very high quality reproduction it is hence necessary to use a light diaphragm stretched until the resonant frequency is at the upper limit of the response range desired, and then to provide just enough damping for the resonant response to be the same as the response at lower frequencies.

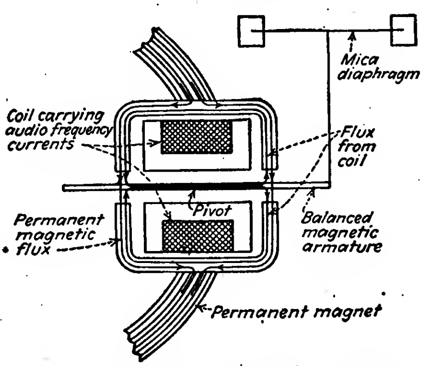


FIG. 463.—Schematic diagram illustrating the operation of a balanced-armature receiver. The forces exerted on the armature are balanced unless the coil is carrying current, in which case the flux produced by the coil unbalances the flux distribution in the air gaps.

Balanced-armature Receivers. The possibilities of the magnetic-diaphragm type of receiver that has just been described are limited by the fact that the diaphragm must be of magnetic material and must withstand the strong steady pull of the permanent magnet. These disadvantages are overcome in the Baldwin or balanced-armature receiver, illustrated in Fig. 463, which employs a pivoted armature arranged as shown in the illustration and connected to a mica diaphragm by means of a link. The pulls exerted on the magnetic armature are balanced until current is passed through the winding, when the additional flux produced by the winding causes an unbalance that deflects the armature.

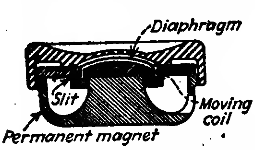


FIG. 464.—Constructional details of a commercial form of moving-coil telephone receiver.

Moving-coil Telephone Receivers.[1]—In the moving-coil type of telephone receiver, the force driving the diaphragm is obtained by the action of a magnetic field on a coil carrying the audio-frequency currents exactly as in the dynamic loud-speaker. The construction of a receiver of this type is shown in Fig. 464. The required damping at resonance is provided by means of a slit which connects the inclosed space back of the diaphragm to outer space, while the inclosed air space behind the diaphragm and the slit provide a coupled system hav-

[1] See E. C. Wente and A. L. Thuras, Moving-coil Telephone Receivers and Microphones, *Bell System Tech. Jour.*, vol. 10, p. 565, October, 1931.

ing mass and stiffness which is utilized to make the response relatively uniform up to frequencies about one and one-half times the resonant frequency of the diaphragm.

155. Microphones.—Any device that converts sound energy into electrical energy is termed a microphone. While many types of microphones have been devised, the only ones that are now used to an appreciable extent are the carbon, condenser, ribbon, crystal, and moving-coil types.

Carbon Microphone.[1]—The carbon microphone makes use of the fact that the resistance which a mass of carbon granules offers to an electrical current depends upon the pressure applied to the carbon. In the carbon microphone a direct current is passed through the granules, which are made from selected and treated anthracite coal and mounted against a diaphragm to form a "button." Sound waves striking the diaphragm vary the pressure exerted on the granules. This produces corresponding changes in resistance of the button and hence causes the current to vary more or less in accordance with the sound pressure exerted against the diaphragm.

The force acting on the diaphragm is proportional to the air pressure of the sound waves, while the pressure upon the granules depends upon the resulting amplitude of diaphragm vibration. As discussed in connection with the magnetic-diaphragm telephone receiver, and shown by Eq. (256), the amplitude of vibration will be substantially independent of frequency below the resonant frequency of the diaphragm, will fall off rapidly above resonance, and at resonance will depend upon the damping. High quality therefore requires a high resonant frequency with suitable damping.

The great advantage of the carbon microphone is that it is an amplifier, since the amount of electrical output that the diaphragm motion controls is greater than the sound power required to operate the diaphragm. At the same time the carbon microphone has a number of disadvantages, the most important of which are the steady background hiss, resulting from random changes in the resistance which the carbon granules offer to the direct current passing through them,[2] and the instability of the carbon "button." As a result the carbon microphone is used in the telephone system where the high sensitivity is a great asset, and

[1] For a more detailed discussion of the carbon microphone, see W. C. Jones, Condenser and Carbon Microphones—Their Construction and Use, *Bell System Tech. Jour.*, vol. 10, p. 46, January, 1931; F. S. Goucher, The Carbon Microphone: An Account of Some Researches Bearing on Its Action, *Bell System Tech. Jour.*, vol. 13, p. 163, April, 1934.

[2] See C. J. Christensen and G. L. Pearson, Spontaneous Resistance Fluctuations in Carbon Microphones and Other Granular Resistances, *Bell System Tech. Jour.*, vol. 15, p. 197, April, 1936.

also in amateur, police, and other radio work where the transmission of information rather than entertainment is the primary object, but is seldom used in public-address or high-quality radio-telephone work.

The construction of a typical high-grade carbon microphone is shown in Fig. 465. The diaphragm is thin duralumin stretched to a resonant frequency of 5700 cycles and associated with a grooved plate which provides the required damping at the resonant frequency and which is provided with two buttons. Carbon microphones used in the telephone system, and also less expensive models, have a single button and obtain increased sensitivity at the expense of uniformity of response by making the diaphragm resonant in the middle of the voice range.

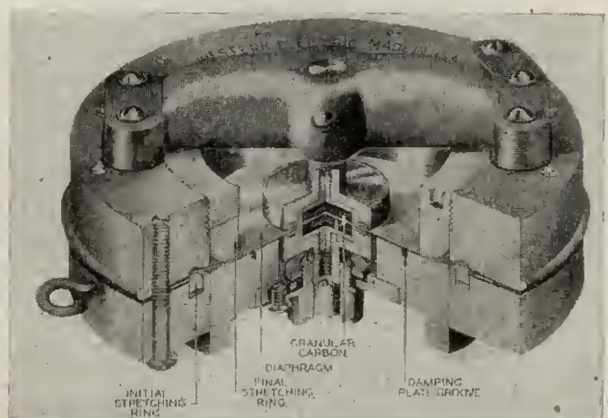


FIG. 465.—Constructional features of high-grade double-button carbon microphone.

Condenser Microphone.[1]—The condenser microphone is a condenser in which one plate is fixed while the other is a diaphragm against which the sound waves act. A direct-current potential of several hundred volts is applied between the plates of the condenser, and, as the capacity is varied by the vibrations that the sound waves produce in the flexible plate, a corresponding voltage drop is produced in the high resistance that is placed between the direct-current voltage and the microphone, as shown in Fig. 467.

The most important constructional details of a typical condenser microphone are shown in Fig. 466. The diaphragm is of aluminum alloy 0.0011 in. in thickness and stretched until its resonant frequency is of the order of 5000 cycles. Acoustic damping is provided by the back plate and is controlled by a series of grooves which intersect each other at right angles, with holes drilled through the back plate at the intersections. The spacing between the diaphragm and back plate must be as

[1] For additional information on condenser microphones see Jones, *loc. cit.*

small as possible in order that movements of the diaphragm will change the capacity appreciably. In the microphone shown the spacing is 0.001 in., and since there is a high direct-current potential difference between the plates it is necessary to keep out all dust and dirt. This is done by sealing the microphone from the outside air.

The condenser microphone is used in the circuit shown in Fig. 467, which should be so proportioned that the capacity shunted across the microphone by the leads and amplifier tube is small in comparison with

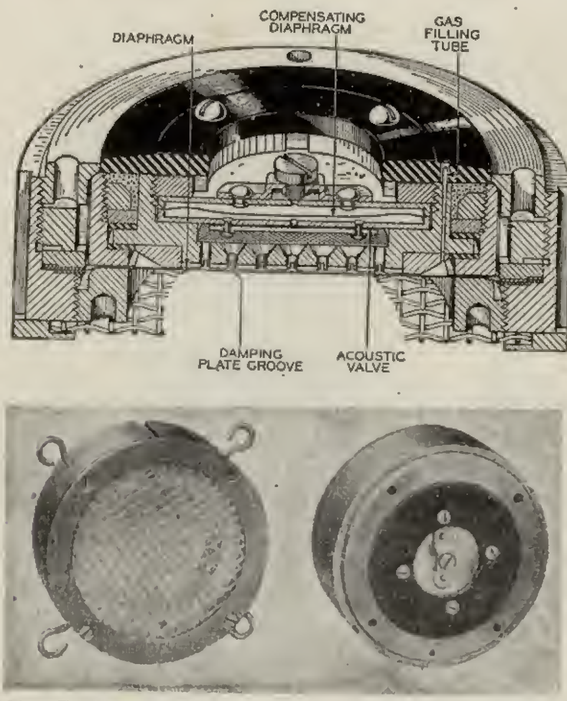


FIG. 466.—Constructional features of a typical condenser microphone.

the microphone capacity, while the equivalent resistance formed by R and R_{gl} in parallel is at least as great as the reactance which the capacity formed by the microphone, its leads, and the amplifier tube has at the lowest frequency to be reproduced. A low shunting capacity increases the sensitivity, since, as the diaphragm vibrates and changes the capacity of the microphone, the resulting potential variations are proportional to the change in capacity divided by the total capacity provided the resistances R and R_{gl} are large enough to prevent appreciable change in the charge on the microphone plates. If these resistances are not large enough, there will be enough charge flowing in and out of the condenser at low frequencies to reduce the potential variations appreciably.

The amplifier tube into which the microphone output feeds should be placed as close as possible to the microphone in order to reduce the capacity of the leads and also to reduce the stray pick-up. This latter point is particularly important because the condenser microphone has a relatively small output and is a high-impedance device. The usual practice is to place the first amplifier tube in the microphone stand. This tube should also be non-microphonic in order to avoid the introduction of extraneous noise through electrode vibration.

The condenser microphone, like the carbon microphone, is a pressure-actuated device in which the output is proportional to the displacement of the diaphragm. The response of a condenser microphone hence is substantially constant at frequencies below diaphragm resonance, falls off rapidly at high frequencies, and has a characteristic at resonance that depends upon the amount of damping present.

A condenser microphone has relatively stable characteristics and can be designed to have substantially constant response over a wide frequency range. Its disadvantages are a relatively low sensitivity and the need of an amplifier close to the microphone. The principal use of the condenser microphone is in making sound measurements. At one time the condenser microphone was widely used in broadcast and public-address work, but it has since been displaced to a considerable extent by the moving-coil, velocity, and crystal microphones.

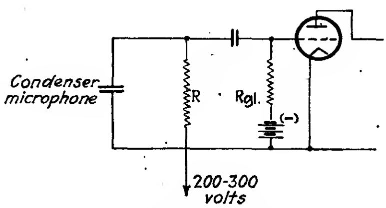


FIG. 467.—Circuit of condenser microphone.

Moving-coil Microphone.[1]—The moving-coil microphone is similar in construction to the moving-coil type of telephone receiver discussed in Sec. 154. The dynamic characteristics of the two instruments differ, however, because in the microphone the voltage induced in the moving coil is proportional to the velocity of the coil, which must therefore be proportional to the sound pressure, while in the telephone receiver the best performance is obtained when the amplitude of vibration varies directly as the pressure of the sound being reproduced. The requirements of the moving-coil microphone can be met by modifying the moving-coil receiver in several respects, and particularly by making the resonant frequency of the diaphragm moderately low, such as 600 cycles, and using resonant air chambers to increase the amplitude at lower frequencies and higher frequencies. When this is properly done, the result is a very excellent microphone having a response practically flat

[1] See E. C. Wente and A. L. Thuras, Moving-coil Telephone Receivers and Microphones, *Bell System Tech. Jour.*, vol. 10, p. 565, October, 1931.

from 40 to 10,000 cycles. In addition to its very satisfactory frequency response the moving-coil microphone has somewhat greater sensitivity than a condenser microphone of the same frequency range and has a low impedance output. This latter feature makes it possible to employ a long twisted-pair cable between microphone and first amplifier without adversely affecting the frequency response and without giving trouble from stray pick-up.

The Ribbon or Velocity Microphone.[1]—The ribbon microphone is a special form of moving-coil microphone in which the moving coil consists of a flat piece of aluminum-alloy foil which is acted upon directly by the sound waves and which has a resonant frequency below the audible range. The construction of such a microphone is shown in Fig. 468.

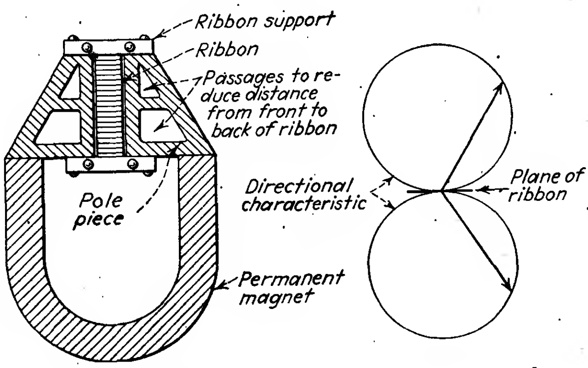


Fig. 468.—Sketch showing constructional details of velocity microphone, together with directional characteristic.

When a plane wave passes by a ribbon microphone, the resulting force acting on the ribbon is proportional to the difference in sound pressure on the front and back of the ribbon. When the frequency is low enough so that the difference in distance to the two sides is appreciably less than a quarter of a wave length, the resulting force exerted against the ribbon is proportional to the frequency and to the pressure gradient, or particle velocity, of the sound wave. As a result, this type of microphone is commonly called a *velocity microphone*. By making the resonant frequency of the ribbon lower than the lowest frequency to be reproduced, the ribbon offers an inertia reactance that is very nearly proportional to frequency. The velocity of vibration and hence the resulting voltage induced in the ribbon are then substantially independent of frequency until the frequency is so great that the difference in distance to the front and to the back of the ribbon approaches a quarter of a wave length, when

[1] See Harry F. Olson, On the Collection of Sound in Reverberant Rooms with Special Reference to the Application of the Ribbon Microphone, *Proc. I.R.E.*, vol. 21, p. 655, May, 1933.

the response tends to fall off. By cutting away the pole pieces as is done in Fig. 468 this difference in distance can be made small enough to give a substantially uniform response to above 10,000 cycles.

When a velocity microphone is placed very close to the sound source, the low frequencies are overemphasized. This comes about because the pressure gradient in a spherical wave is greater in proportion to pressure than the pressure gradient of a plane wave by the factor $\sqrt{1 + (c/\omega r)^2}$, where r is the distance to the source, c is the velocity of sound, and ω is 2π times frequency.

The velocity microphone also has a pronounced directional characteristic as shown in Fig. 468. This is because sound waves arriving from the side strike both the front and back of the ribbon at the same instant and so produce no resultant force. This directivity can often be taken advantage of to minimize undesired reverberation and noises, and also makes the velocity microphone particularly suitable for use as a lapel microphone, since by properly utilizing the directional characteristic it is possible to arrange matters so that the output of a microphone located upon the coat lapel is substantially independent of the direction in which the speaker's head is turned.[1]

The velocity microphone has the advantage of simplicity, an extremely good frequency response, and a low output impedance. Its principal disadvantage is that it cannot be used close to the source of sound.

Crystal Microphone.[2]—Crystal microphones make use of the piezoelectric effect of Rochelle salt crystals to transform mechanical stress produced by sound waves into electrical output. The most successful arrangement of this type consists of two crystal slabs mounted as shown in Fig. 469. The bending of these slabs resulting from the pressure of the sound wave produces a voltage across the crystal face that is then utilized to operate an amplifier. By making the assembly small so that the resonant frequency is above the audible range, the device becomes pressure operated, and with simple equalizing circuits gives a substantially uniform response up to 15,000 cycles. The small size of the crystal assembly also makes this type of microphone practically non-directional. The electrical output of a single unit such as illustrated in Fig. 469 is small, but can be readily increased by connecting a number of these in series.

[1] Harry F. Olson and Richard W. Carlisle, A Lapel Microphone of the Velocity Type, *Proc. I.R.E.*, vol. 22, p. 1354, December, 1934.

[2] See Stuart Ballantine, High Quality Radio Broadcast Transmission and Reception, *Proc. I.R.E.*, vol. 22, p. 564, May, 1934; C. B. Sawyer, The Use of Rochelle Salt Crystals for Electrical Reproducers and Microphones, *Proc. I.R.E.*, vol. 19, p. 2020, November, 1931.

The crystal microphone is simple and has a good frequency response. Its chief disadvantages are the relatively low sensitivity, the fact that the sensitivity tends to fall off at temperatures exceeding 90°F., and high output impedance.

Some use has been made of crystal microphones in which sound waves act upon a diaphragm that is mechanically coupled in such a way as to produce a stress in the crystal. Such microphones have relatively high sensitivity and are inexpensive, but do not generally have a very uniform frequency response.

Pressure- and Field-response Characteristics.—Condenser, moving-coil, and similar microphones give responses that depend upon the pressure against the diaphragm, and are normally designed so that over the essential frequency range the output is proportional to pressure and independent of frequency. The calibration curve of such a microphone depends, however, upon the way in which the calibration is made because the pres-

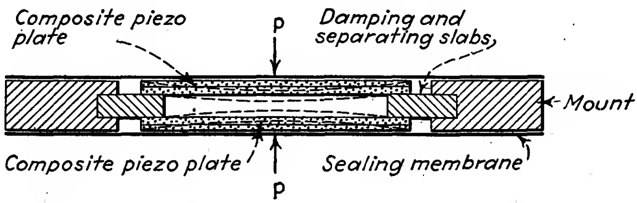


FIG. 469.—Construction of "sound-cell" of crystal microphone.

sure of a sound wave traveling in free space is not necessarily the pressure that this same wave produces acting against the diaphragm of a microphone. At low frequencies, where the microphone dimensions are small compared with a wave length, the sound wave diffracts around the microphone with negligible reflection and produces a pressure against the diaphragm which is a true measure of the pressure of the wave. At higher frequencies, however, reflections occur at the microphone and may cause the pressure exerted on the diaphragm to reach twice the pressure of the wave. In addition, the front of the microphone forms a shallow air pocket which introduces a resonance that will still further increase the pressure against the diaphragm at high frequencies.

The relation between the output voltage of a condenser microphone and the pressure exerted against the diaphragm is called the *pressure calibration*, while the relation between the pressure of a free sound wave and the output voltage which this wave develops when striking the microphone is called the *field calibration*. The two are the same at low frequencies, while at high frequencies the field calibration depends upon the angle of incidence with which the wave strikes the microphone, as well

as upon the frequency. Typical pressure- and field-calibration curves of a condenser microphone are shown in Fig. 470.[1]

156. Measurement of Microphone and Loud-speaker Characteristics. Sound measurements are ordinarily made with the aid of a calibrated microphone, most commonly a condenser microphone. In making measurements, the difference between the field and pressure calibrations is a source of difficulty if accurate results are required. This trouble can be eliminated, however, by placing the condenser microphone in a spherical housing and employing a construction that eliminates the cavity.[2] The relationship between field and pressure calibrations

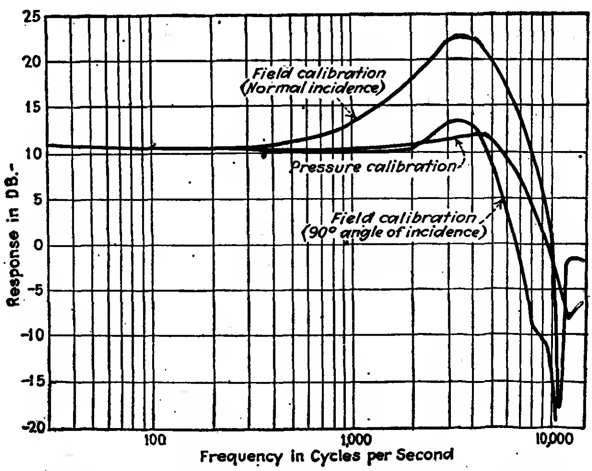


FIG. 470.—Typical pressure and field calibrations of a condenser microphone.

can then be calculated because of the simple geometrical shape. Another expedient sometimes employed consists in using a miniature condenser microphone so small that the field and pressure calibrations do not differ until above the useful range of frequencies.[3] Such miniature microphones are also often used in exploring sound fields in horns, etc.

Microphone Calibration.[4]—The field calibration of a microphone can be most conveniently obtained by using a Rayleigh disk to measure the

[1] For a more detailed discussion of field and pressure calibrations, see L. J. Sivian, Absolute Calibration of Condenser Transmitters, *Bell System Tech. Jour.*, vol. 10, p. 96, January, 1931.

[2] See Stuart Ballantine, Note on the Effect of Reflection by the Microphone in Sound Measurements, *Proc. I.R.E.*, vol. 16, p. 1639, December, 1928.

[3] H. C. Harrison and P. B. Flanders, An Efficient Miniature Condenser Microphone System, *Bell System Tech. Jour.*, vol. 11, p. 451, July, 1932.

[4] Also see The Calibration of Microphones, *Elec. Eng.*, vol. 55, p. 241, March, 1936.

intensity of the free wave.[1] The Rayleigh disk consists of a very light silvered disk hung vertically on a fine suspension and arranged with an optical system so that small rotations can be accurately measured by means of an optical lever. Sound waves in striking the disk tend to turn it perpendicular to the direction of travel of the wave by an amount proportional to the pressure gradient (or particle velocity) of the sound wave.

Absolute pressure calibrations can be made by means of the thermophone, the Rayleigh disk, or an actuator.[2] The thermophone method produces calculable sound pressures in a small inclosed space in the front of the microphone by passing an alternating current superimposed upon a direct current through a strip of gold or platinum foil so thin that the temperature varies at the frequency of the alternating current. In the Rayleigh-disk method the microphone is placed at the closed end of a resonance tube, and the Rayleigh disk is used to measure the velocity at a pressure node.

In the actuator method a calculable force is applied to the diaphragm by electrostatic means. Such methods are primarily applicable to the condenser microphone, and can be carried out in a number of ways. A typical arrangement consists in subjecting the microphone to sound waves and then preventing the diaphragm from vibrating by means of an equal and opposite force. This balancing force can be obtained by applying between the diaphragm and back plate a voltage of the same frequency as the sound and of adjustable magnitude and phase. The magnitude of the resulting electrostatic attraction at balance can be calculated from the dimensions involved, and equals the pressure of the sound wave upon the diaphragm. The condition of balance can be detected by making the capacity between diaphragm and back plate a part of an oscillating circuit and by adjusting the balancing force until there is no frequency modulation.

Loud-speaker Characteristics.[3]—The frequency-response characteristic of a loud-speaker is normally obtained by exciting the speaker under

[1] An excellent description of the details involved in such a calibration is given by Harry F. Olsen and Stanford Goldman, The Calibration of Microphones, *Electronics*, vol. 3, p. 106, September, 1931.

[2] For a more detailed discussion of pressure calibrations, see L. J. Sivian, Absolute Calibration of Condenser Transmitters, *Bell System Tech. Jour.*, vol. 10, p. 96, January, 1931.

[3] A summary of factors involved in loud-speaker measurements is given by Frank Massa, Loud Speaker Measurements, *Electronics*, vol. 9, p. 18, July, 1936. Also see L. G. Bostwick, Acoustic Considerations Involved in Steady State Loud-speaker Measurements, *Bell System Tech. Jour.*, vol. 8, p. 135, January, 1929; I. Wolff and A. Ringel, Loud-speaker Testing Methods, *Proc. I.R.E.*, vol. 15, p. 363, May, 1927; I. Wolff, Sound Measurements and Loud-speaker Characteristics, *Proc. I.R.E.*, vol.

normal conditions and observing the sound with the aid of a calibrated microphone. In carrying out such measurements it is desirable that as much as possible of the sound reaching the microphone come directly from the speaker instead of indirectly as a result of a reflection. Otherwise interference effects will modify the pressure produced at the microphone in a relatively complicated manner. Loud-speaker characteristics are accordingly obtained most satisfactorily from outdoor measurements where the speaker can be mounted on a high stand or hung from ropes in such a manner as to be remote from reflecting surfaces. Where characteristics must be taken indoors, as is necessary in routine testing, this is best done in a room lined with sufficient sound-absorbing material to minimize the reflected waves. Even then, the reflections at low frequencies, where the sound-absorbing material is least effective, will be sufficient to produce pronounced standing wave patterns that will be very critical with respect to frequency. The resulting irregularities in the response curve can be ironed out by rotating the microphone to get an indication proportional to the average sound pressure over a considerable area, although the presence of reflections still increases the apparent response observed. This effect is most pronounced at low frequencies, where sound absorbents are least effective, and results in indoor measurements always indicating a greater low-frequency output than actually exists.

In the measurement of loud-speaker characteristics the effect of reflections can be minimized by placing the microphone as close as possible to the loud-speaker in order to increase the ratio of direct to indirect sound intensities. There is a limit to the extent that one can go in this direction, however, since with the microphone very close to a loud-speaker the low frequencies register stronger than they really are.

In testing the over-all characteristics of radio receivers it is necessary to take into account the acoustics of the space (commonly the home living room) in which the receiver reproduces its sound. Considerable attention has been given to this problem in connection with the development of high-fidelity receivers, and the factors that must be considered are dicussed at length in the literature.[1]

The efficiency of a loud-speaker can be determined by measuring the electrical input to the voice coil and observing the total sound power

16, p. 1729, December, 1928; Benjamin Olney, Notes on Loud-speaker Response Measurements and Some Typical Response Curves, *Proc. I.R.E.*, vol. 19, p. 1113, July, 1931.

[1] Harold A. Wheeler and Vernon E. Whitman, Acoustic Testing of High Fidelity Receivers, *Proc. I.R.E.*, vol. 23, p. 610, June, 1935; Stuart Ballantine, High Quality Radio Broadcast Transmission and Reception, *Proc. I.R.E.*, vol. 23, p. 618, June, 1935.

produced. The total power can be determined by measuring the sound intensity in different directions and then integrating the resulting energy over a spherical surface. This, however, requires an enormous amount of labor if an accurate evaluation of efficiency is to be obtained over the entire frequency range An alternative procedure for measuring total sound output consists in operating the loud-speaker in a room having a high reverberation time. In such a room the sound pressure at every point tends to be the same and has a value determined by sound power, reverberation time, and the room volume. If precautions are taken to cause a uniform distribution by warbling the frequency of the exciting power or by using a large rotating paddle to reduce standing waves, rather accurate results can be obtained quite easily.

A knowledge of the motional impedance of a loud-speaker makes it possible to estimate the sound output at least approximately, as well as to obtain other useful information concerning the details of speaker oper-

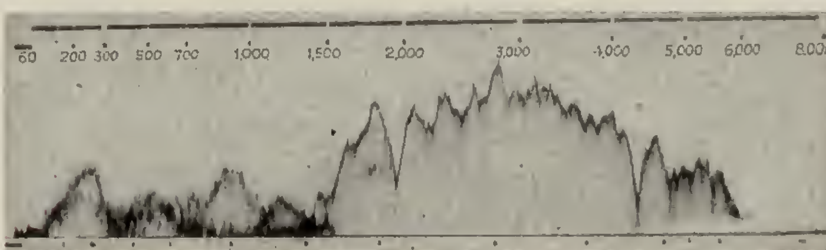


FIG. 471.—Loud-speaker characteristic as recorded on an oscillograph film.

ation. The motional impedance can be obtained by taking the difference between the impedance of the voice coil under normal operating conditions and the impedance when the voice coil is prevented from vibrating. The resistance component of the resulting motional impedance accounts for the radiation resistance, eddy-current loss, and various forms of frictional loss in cone and spider. The latter losses can be separately evaluated by measuring the motional impedance of the same speaker when mounted in a bell jar from which the air has been exhausted. The difference between the motional resistance in a vacuum and under normal operation is largely accounted for by the radiation resistance. Upon this assumption the fraction of the input electrical energy that is dissipated in the radiation-resistance part of the motional resistance can be readily calculated. This method, while relatively simple, is not highly accurate with paper cones unless the resistance component of the motional impedance is reasonably large, because the diaphragm loading is not the same in a vacuum as in air, and the cone losses will accordingly be different.

The response of most loud-speakers contains so many irregularities that a point-by-point calibration taken at regular frequency intervals

is extremely tedious. A more satisfactory method of testing a loud-speaker is to drive it with a beat-frequency oscillator, hunt for the peaks and valleys in the characteristic, and make observations at these points irrespective of the exact frequency intervals that result. Even with such a procedure the time required to obtain a complete calibration is very great. When a large number of characteristics are to be obtained, a more satisfactory test procedure is to employ a photographic or curve-tracing system of recording. A typical example consists of a two-element oscillograph arranged so that one element records the output level of the microphone amplifier. The other is associated with the dial of the beat-frequency oscillator in such a way as to give indications at certain known frequencies by interconnecting the beat-frequency oscillator and oscillograph drum mechanically. It is possible in this way to take a complete calibration in a fraction of a minute. An example of such a calibration is shown in Fig. 471.[1]

Problems

1. Explain how it is that men and women are able to produce the same vowel sounds in spite of the differences in pitch of male and female voices.

2. Explain why it is found that musical instruments producing low notes must develop much more power to be effective than when pitches of the order of 500 to 1000 cycles are involved.

3. Explain why subjective tones are much more pronounced when dealing with loud sounds than with weak sounds.

4. Explain why high-pitched sounds mask low frequencies only slightly if at all.

5. A particular piece of music when listened to in a theater is assumed to have a loudness level of 80 db according to Fig. 448. If reproduced with a loudness level of 50 db in the home by a radio, calculate and plot with the aid of Fig. 448 the relative response as a function of frequency which the compensated tone control must provide to make up for the difference in reproduction level.

6. Explain why the optimum reverberation time for an auditorium is different when a speaker is addressing a large audience with and without the help of a public-address system.

7. In view of the situation illustrated in Fig. 448, explain why the reverberation time should be different for different frequencies.

8. Explain why sounds such as *s*, *f*, *ch*, *v*, and *sh* are reproduced very poorly, if at all, by the ordinary telephone, while the vowel sounds are always understandable.

9. In the transatlantic telephone it is found desirable deliberately to introduce frequency distortion by increasing the relative intensity of the voice frequencies above 1000 cycles before transmission, and then after reception to restore the original proportions. What advantages can be expected from such an arrangement?

10. Why is it that articulation tests are never used in testing broadcast stations, but are employed extensively in connection with radio extensions of the telephone system?

11. Demonstrate with the aid of Eqs. (253) to (258) that at frequencies high enough for the radiation resistance to be independent of frequency the total sound

[1] This particular technique for the rapid recording of loud-speaker characteristics was developed in 1930 by C. R. Skinner, graduate student at Stanford University.

power radiated for a given voice-coil current is greater the smaller the effective area of the diaphragm if it is assumed that the effective mass is proportional to the effective area.

12. Calculate the required size of baffle as a function of the lowest frequency to be reproduced over the range 50 to 250 cycles, and from the results discuss: (*a*) the necessity of inclosing the back side of the loud-speaker or of using some form of acoustic labyrinth or network if frequencies below 125 cycles are to be satisfactorily reproduced by a loud-speaker mounted in an ordinary radio-receiver cabinet, (*b*) the low-frequency response that can be expected of midget radio receivers.

13. Demonstrate with the aid of Eqs. (253) to (258) and Fig. 454 that, at low frequencies where the cone diameter is appreciably less than a half wave length and piston action can be expected, the efficiency increases as the cone diameter is increased. Assume that the mass of the moving system and the length of the voice-coil wire (and hence the voice-coil resistance) are proportional to the area of the cone, and be sure to take into account the change in input power with change in voice-coil resistance.

14. Derive Eq. (258).

15. In an 8-in. cone speaker the coil and cone mass is 14 grams, the compliance of the suspension 5×10^{-7} cm per dyne, the length of wire in voice coil 1200 cm, and the air-gap flux density 8000 gausses. Assuming piston action, and assuming a voice coil current of one ampere, calculate and plot (*a*) force on voice coil, (*b*) total radiation resistance for both sides of cone, (*c*) mechanical impedance of vibrating system, (*d*) velocity of vibration system, (*e*) total sound power radiated, and (*f*) motional impedance, as a function of frequency over the range 80 to 1000 cycles. Neglect the mass effect of the fluid air, and also all losses in the vibrating system except radiation resistance.

16. Determine the rate of taper, the mouth area, throat area, and length of a horn which will reproduce frequencies down to 100 cycles and which will load a diaphragm having an effective area of 100 sq cm with an acoustic load of 300 mechanical ohms per square centimeter.

17. Explain why the mass of diaphragm and coil in a horn speaker causes the response to fall off at very high frequencies but does not have much effect at low and moderately low frequencies.

18. What would be the effect upon the sound output of the horn in Fig. 460 if the sound-absorbing material were omitted from the box inclosing the back side of the cone?

19. When a loud-speaker system employing horns is called upon to develop a very large power output, explain why a dual speaker system (separate high- and low-frequency units) will eliminate substantially all trouble from non-linear action of the air in the throat of the horn.

20. Explain why the quality of sound output developed by a telephone receiver will change when the receiver is removed from the ear.

21. In pressure-operated microphones such as the carbon and condenser types, the output voltage for a given sound pressure decreases as the resonant frequency to which the diaphragm is stretched is increased. Explain.

22. In the velocity microphone, increasing the distance from front to back of the ribbon increases the sensitivity in direct proportion to the reduction in high-frequency limit. Explain the reason for this.

23. Calculate and plot as a function of frequency the relative response of a velocity microphone when the distance from the microphone to sound source is 1, 2, and 5 ft. Cover the frequency range 50 to 500 cycles.

APPENDIX A

FORMULAS FOR CALCULATING INDUCTANCE, MUTUAL INDUCTANCE, AND CAPACITY

157. Formulas for Calculating Inductance, Mutual Inductance, and Capacity.[1]—This section gives formulas for calculating inductance, mutual inductance, and capacities for the cases commonly encountered in radio work. Most of these formulas involve small approximations but will give results sufficiently accurate for all ordinary requirements.

Inductance of Single-layer Solenoid.—The equation applying in this case has already been discussed in Sec. 6 and is

$$\text{Inductance in microhenries} = n^2dF \tag{262}$$

where n is the number of turns, d is the diameter of the coil measured to the center of the wire, and F is a constant determined by the ratio of length to diameter, and given in Fig. 7. This formula can also be used to obtain the inductance of multilayer coils provided that the radial depth of the winding is small compared with the radius and length of the coil. In such cases the equivalent diameter is taken as the diameter measured to the center of the winding. Equation (262) can also be used to calculate the inductance of polygonal coils, when the number of the sides of the polygon is fairly large, by assuming that the coil is equivalent to a helix whose mean radius is the mean of the radii of the circumscribed and inscribed circles of the polygon.

Inductance of Single-layer Rectangular Coil.

$$\text{Inductance in microhenries} = an^2[G + H] \tag{263}$$

where

a = length of long side in inches
n = number of turns
G = factor determined by a_1/a and b/a and given in Fig. 472
H = factor determined by number of turns and $\dfrac{\text{diameter of wire}}{\text{length of coil}}$ and given in Fig. 472.
a_1 = length of short side in inches
b = axial length of coil in inches

[1] Most of the formulas given here are taken from *Bur. Standards Circ.* 74, Radio Instruments and Measurements. This book is the standard authority on the subject and contains formulas for making calculations of any desired accuracy for almost every case that can be encountered in practice.

This formula finds its chief application in calculating the inductance of loop antennas.

Inductance of Multilayer Coils with Winding Having a Rectangular Cross Section.

$$\text{Inductance in microhenries} = \frac{l^{5/3}}{D^{2/3}} I \tag{264}$$

or

$$\text{Inductance in microhenries} = an^2 J \tag{265}$$

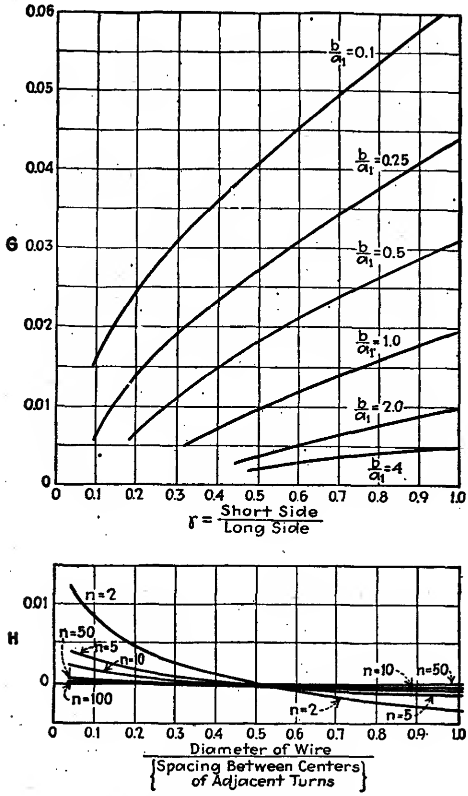


FIG. 472.—Factors for use in Eqs. (263) and (266).

where l is the length of wire in inches, D the distance between centers of adjacent wires, a the mean coil radius in inches, n the number of turns, and I and J factors given by Fig. 473. It will be observed that the maximum inductance is obtained from a given length of wire when the cross section of the winding is square and the side of the cross section is 0.662 times the mean coil radius.

Inductance of a Flat Spiral.—A flat spiral is a special case of a multilayer coil having a winding of rectangular cross section, and so can be handled by Eq. (265). With a fixed width of coil, the maximum inductance with a given length of wire is obtained when the radial depth of the coil is three-fourths of the mean radius.

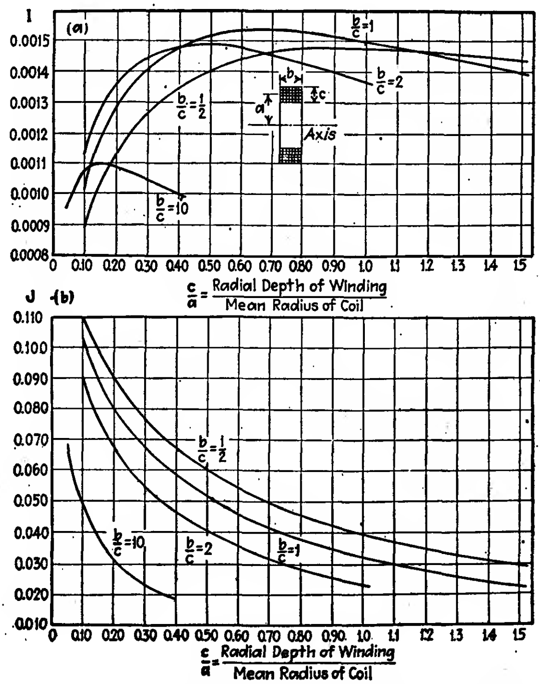


FIG. 473.—Factors for use in Eqs. (264) and (265).

Inductance of Multilayer Rectangular Coils Having Windings of Rectangular Cross Section.

$$\text{Inductance in microhenries} = an^2G. \tag{266}$$

where a is the length in inches of the longer side of the rectangle measured between centers of the rectangular winding cross section, n is the number of turns, and G is given by Fig. 472, where a/a_1 is taken as the ratio of short side to long side of the rectangle and b/a is considered to be

$$\frac{\text{circumference of winding cross section}}{2a}$$

Inductance of Toroidal Coils with Single-layer Windings.—When the toroid is a circular ring of circular cross section (doughnut shape)

$$\text{Inductance in microhenries} = 0.0319n^2R\left[1 - \sqrt{1 - \left(\frac{a}{R}\right)^2}\right] \quad (267)$$

where R is the distance in inches from the axis to center of cross section of winding, a is the radius of the turns of the winding, and n is the number of turns.

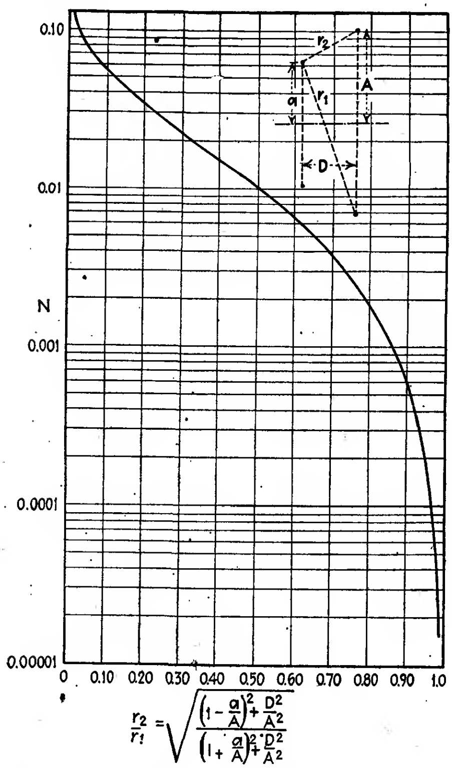


FIG. 474.—Factor for use in Eq. (271).

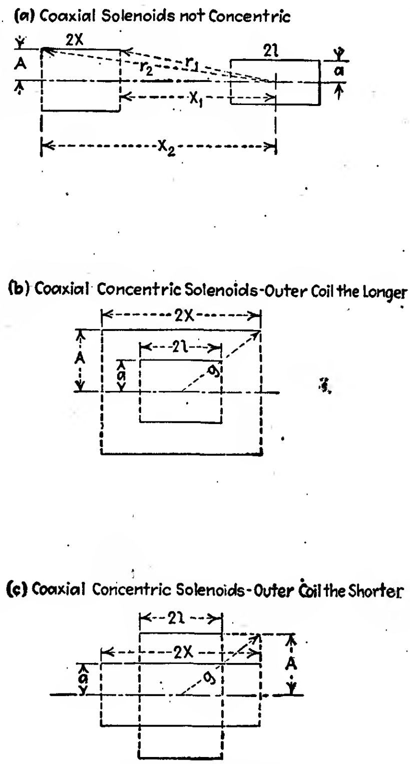


FIG. 475.—Possible arrangements of coaxial solenoids.

When the toroid is a ring of rectangular cross section

$$\text{Inductance in microhenries} = 0.0117n^2h \log_{10} \frac{r_2}{r_1} \quad (268)$$

where n is the number of turns, h the axial length of the ring in inches, and r_1 and r_2 the inner and outer radii of the ring.

Inductance of Parallel-wire Transmission Line.

$$\text{Inductance in microhenries per foot} = 0.281 \log_{10} \frac{b}{a} + 0.030 \quad (269)$$

where b is the spacing between centers and a the radius of the wire.

Inductance of Cable with Sheath Serving as Return Conductor.

$$\text{Inductance in microhenries per foot} = 0.140 \log_{10} \frac{b}{a} + 0.015 \qquad (270)$$

where b is the radius of the inner side of the sheath and a is the radius of the central conductor. This formula neglects the small contribution to the inductance made by the flux within the sheath metal.

Mutual Inductance between Two Parallel Coaxial Circles.—Using the notation shown in Fig. 474, with the dimensions in inches

$$\text{Mutual inductance in microhenries} = N\sqrt{Aa} \qquad (271)$$

where N depends upon r_2/r_1 and is given by Fig. 474.

Mutual Inductance between Two Coaxial Circular Coils of Rectangular Cross Section.—If the windings are of square or approximately square cross section, then

$$\text{Mutual inductance in microhenries} = n_1 n_2 M_0 \qquad (272)$$

where n_1 and n_2 are the number of turns in the two coils, and M_0 is the mutual inductance as calculated by Eq. (271) for two coaxial circles which are located at the centers of the cross sections of the two coils. Equation (272) will hold with good accuracy even when the cross section departs widely from a square provided the coils are not close together.

Mutual Inductance between Coaxial Solenoids.—There are three cases to distinguish, as illustrated in Fig. 475.

Coaxial solenoids not concentric

$$M = 0.02505 \frac{a^2 A^2 n_1 n_2}{2x \times 2l} [K_1 k_1 + K_3 k_3 + K_5 k_5] \qquad (273)$$

where

a = the smaller radius, measured from the axis of the coil to the center of the wire, in inches

A = the larger radius, measured in the same way, in inches

$2l$ = length of the coil of smaller radius = number of turns times the pitch of winding, in inches

$2x$ = length of the coil of larger radius, measured in the same way, in inches

n_1 and n_2 = total number of turns on the two coils

$$K_1 = \frac{2}{A^2}\left(\frac{x_2}{r_2} - \frac{x_1}{r_1}\right), \qquad k_1 = 2l$$

$$x_1 = D - x \qquad r_1 = \sqrt{x_1^2 + A^2}$$

$$x_2 = D + x \qquad r_2 = \sqrt{x_2^2 + A^2}$$

D = axial distance between centers of the coils in inches.

$$K_3 = \frac{1}{2}\left(\frac{x_1}{r_1^5} - \frac{x_2}{r_2^5}\right), \qquad k_3 = a^2 l\left(3 - 4\frac{l^2}{a^2}\right)$$

$$K_5 = -\frac{A^2}{8}\left[\frac{x_1}{r_1^9}\left(3 - 4\frac{x_1^2}{A^2}\right) - \frac{x_2}{r_2^9}\left(3 - 4\frac{x_2^2}{A^2}\right)\right]$$

$$k_5 = a^4 l\left(\frac{5}{2} - 10\frac{l^2}{a^2} + 4\frac{l^4}{a^4}\right)$$

Concentric Solenoids (Figs. 475*b* and 475*c*).—The formulas for these two cases are the same provided g is defined for each case as shown in the figure, and the dimensions are in inches.

$$\text{Mutual inductance in microhenries} = 0.0501\frac{a^2 n_1 n_2}{g}\left[1 + \frac{A^2 a^2}{8g^4}\left(3 - 4\frac{x^2}{a^2}\right)\right] \qquad (274)$$

Capacity of Parallel-plate Condenser.

$$\text{Capacity in micromicrofarads} = 0.2244K\frac{S}{t} \qquad (275)$$

where

K = dielectric constant

S = area of dielectric in square inches

t = thickness of dielectric in inches.

Capacity of Two-wire Transmission Line.

$$\text{Capacity in micromicrofarads per foot} = \frac{3.680}{\log_{10}(b/a)} \qquad (276)$$

where b is the spacing between wire centers and a is the radius of the wire.

Capacity of Round Wire in a Concentric Sheath.

$$\text{Capacity in micromicrofarads per foot} = \frac{7.354K}{\log_{10}(b/a)} \qquad (277)$$

where K is the dielectric constant of the insulation ($K = 1$ for air), b is the radius of the inner side of the sheath, and a is the radius of the central conductor.

NAME INDEX

D

E

F

G

H

N

O

P

Q

R

S

SUBJECT INDEX

B

Q

R

RADIO ENGINEERING

TERMAN

McGRAW-HILL BOOK COMPANY