THE BELL SYSTEM
TECHNICAL JOURNAL

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

TABLE OF CONTENTS
AND
INDEX
VOLUME XXV
1946

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

PRINTED IN U. S. A

THE BELL SYSTEM

TECHNICAL JOURNAL

VOLUME XXV, 1946

Table of Contents

OCTOBER, 1946

Index to Volume XXV

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

Because circumstances have delayed the issuance of this number of the Bell System Technical Journal far beyond its designated date of appearance, it seems desirable to record that issuance actually occurred on June 10, 1946.

Page 331: Equation

$$y_0(t) = y_0(t_0) + \frac{\dddot{y}_0(t_0)[ut_1 + u^2 t_2 + u^3 t_3 + \cdots]}{1!}$$

$$+ \frac{\dddot{y}_0(t_0)[ut_1 + u^2 t_2 + u^3 t_3 + \cdots]^2}{2!} + \cdots$$

should read

$$y_0(t) = y_0(t_0) + \frac{\dot{y}_0(t_0)[ut_1 + u^2 t_2 + u^3 t_3 + \cdots]}{1!}$$

$$+ \frac{\ddot{y}_0(t_0)[ut_1 + u^2 t_2 + u^3 t_3 + \cdots]^2}{2!} + \cdots$$

Page 332: Equation $t_1 = -\dfrac{y_1'(t_c)}{\dddot{y}_0(t_0)}$

should read $t_1 = -\dfrac{y_1(t_0)}{\ddot{y}_0(t_0)}$

In last part of first paragraph

expression $(\dot{y}^2)y = 0$

should read $(\dot{y}^2)_{y=a}$

Equation $K_1 = 2(\dddot{y}_0 \ddot{y} - y_0 \dddot{y}_1)$

should read $K_1 = 2(\dot{y}_0 \dot{y}_1 - \ddot{y}_0 y_1)$

Equation $K_2 = (\dddot{y}_1^2 - 2y_1 \dddot{y}_1 + 2\dddot{y}_0 \dddot{y}_2) - 2\dddot{y}_0 y_2 + \dfrac{\dddot{y}_0 y_1^2}{\dddot{y}_0}$

should read $K_2 = (\dot{y}_1^2 - 2y_1 \dddot{y}_1 + 2\dot{y}_0 \dot{y}_2) - 2\ddot{y}_0 y_2 + \dfrac{\dddot{y}_0 y_1^2}{\dot{y}_0}$

Page 333: Expression $\cos\left(\dfrac{\pi y}{b} + c\right) = \cos\left(\dfrac{\pi y_0}{b} + c\right)$

$$+ \frac{\dfrac{\pi}{b}\sin\left(\dfrac{\pi y_0}{b} + c\right)[uy_1 + u^2 y_2 + \cdots] + \cdots}{1!}$$

should read $\cos\left(\dfrac{\pi y}{b} + c\right) = \cos\left(\dfrac{\pi y_0}{b} + c\right)$

$$- \frac{\dfrac{\pi}{b}\sin\left(\dfrac{\pi y_0}{b} + c\right)[uy_1 + u^2 y_2 + \cdots] + \cdots}{1!}$$

Early Fire-Control Radars for Naval Vessels

By W. C. TINUS and W. H. C. HIGGINS

Introduction

FOR a number of years before the war a very intensive development effort was under way in the Army and Navy laboratories, and in several commercial laboratories, on the application of radio methods to the location of objects at a distance. The equipment which resulted was eventually called "Radar" equipment by the Navy and this term is now almost universally used. The urgent needs of the war have resulted in the very rapid development and extensive application of this new science during the last few years.

Radar equipments of many different types have been designed to perform specific functions on land and sea, and in the air. These equipments have had an important part in the winning of the war and the recent relaxation in secrecy regulations now permits publishing some of the story. In this present article a description of the Mark 3 and 4 Fire-Control Radars for Naval Vessels will be given, together with a little of the history that preceded their development.

Historical Background

When the Bell Telephone Laboratories began active radar development work early in 1938 an effort was made to set technical objectives for this work that would avoid duplication of the intensive work then under way in the Army and Navy laboratories, and that would advance the art toward the solution of some of the recognized basic problems. The general objectives were to increase the accuracy of radar measurement of location and to increase as much as possible the operating carrier frequency. The reasons for these objectives are discussed in the following paragraphs.

The state of the art at the time under discussion has been partially described in a recent paper by Maj. Gen. R. B. Colton.[1] The work he described and directed was carried out at the Signal Corps Laboratories at Fort Monmouth, New Jersey and was directed principally toward solving

---

[1] "Radar in the U. S. Army" by Maj. Gen. Roger B. Colton, published in the *Proceedings of the I. R. E.*, November, 1945.

the ground forces' problems of aircraft warning and searchlight control. At the same time intensive work was being pursued at the Naval Research Laboratory at Anacostia, D. C. under the direction of Dr. A. H. Taylor, Dr. R. M. Page and Mr. L. C. Young. Their work was directed primarily toward developing radar equipment that would be useful aboard ship, and it was from them and from the engineers of the Navy Department that the principal inspiration and guidance for the work described in this paper were obtained.

The first military application in which radar equipment proved its usefulness was in the detection of approaching aircraft. For this kind of application the radar is not required to locate the approaching planes with very great accuracy and the experimental radars of 1938 and 1939 performed this function in quite a useful way. The fact that the first application of radar was a strictly defensive one may account in part for the great interest and support given radar work in England and in this country, while apparently much less radar work was done before the war by the scientists of Germany and Japan. Thus, when radar later became a powerful and versatile aid to offense, the enemy nations found themselves years behind in development.

Very early in their work the men of the Naval Research Laboratory recognized the potential ability of radar to help solve the fire-control problem. Since this problem determined the design of the radar systems to be described later in this paper a brief general discussion of fire control is given here. The term *fire control* refers broadly to the means by which a gun or other weapon is aimed and fused so that, when fired, the projectile will hit or burst near the intended target. A fire-control system includes two major parts: first, a locating device for determining the present position of the target; and second, a computing device which analyzes the present position data, computes the target's course and speed, and the position the target will occupy at the future time when the projectile arrives at that point, and finally furnishes the correct aiming and fusing information to the guns. A modern fire-control system does these things in a continuous manner so the guns remain correctly aimed and can be fired at any time during the engagement.

Before the war the present position of the target was ordinarily determined by optical instruments. Operators tracked the target by controlling their telescopes in such a way that the target remained on the crosshairs in their eyepieces. Thus the azimuth and elevation angles were found. Another operator measured the range to the target with an optical range finder, or indirectly estimated range from the angular extent of the target and its estimated size.

The accuracy of this optical system in determining azimuth and elevation

angles is very good provided the target can be seen clearly. This proviso is a serious limitation under many typical operating conditions. It is frequently difficult to see a target at a range of several miles on account of haze even on a relatively clear day, and at night or in fog or smoke screen the usefulness of a telescope is almost nil. The optical range finder is subject to the same limitations as the telescope and in addition leaves much to be desired in the matter of accuracy and continuity of data even under the best visibility conditions. This is due to the fact that optical range finders are triangulation devices which inherently have accuracy limitations. The need for a long and very stable base line between the prisms of an optical range finder is difficult to meet aboard ship, and the principle of operation makes inevitable a rapidly decreasing accuracy with increasing range. Thus, as the effective range of guns increased, the need for more accurate means for measuring range became more acute.

In its earliest forms radar offered at once a potential means for measuring range with much better accuracy than that of the optical range finder. This was due to the different principle on which radar works. A pulse of radio frequency energy is sent out to the target and the echo signal is received back at the source. The velocity of the waves en route is the same as that of light, and is one of the basic physical constants. To measure range accurately with radar required only the development of techniques for producing short transmitted pulses and for measuring accurately the short intervals of time between the transmitted pulse and the returning echo pulse. Both of these were the kind of problems which yield readily to electronic solutions. The early work in Bell Telephone Laboratories thus included the production of shorter transmitted pulses than were being commonly used, and the development of improved range measuring means.

The second important general objective for the early work at Bell Telephone Laboratories was to devise equipment which would operate at frequencies much higher than had been previously used. The need for higher-frequency operation arose from the fact that for a given size of antenna the beam width decreases with increasing frequency while the gain increases. Narrow beams are required to obtain accurate angular data while increased gain is desirable since it obviously provides increased range for a given transmitter power and receiver noise figure. These factors are illustrated by the curves of Figs. 1A and 1B which show the relationship between beam width, antenna gain and antenna size expressed in wavelengths. The curve labeled "uniform illumination" yields maximum gain and minimum beam width for a given antenna size but produces unwanted side lobes of undesirable amplitude. For this reason the illumination is usually graded over the antenna aperture to reduce minor lobes. The gain
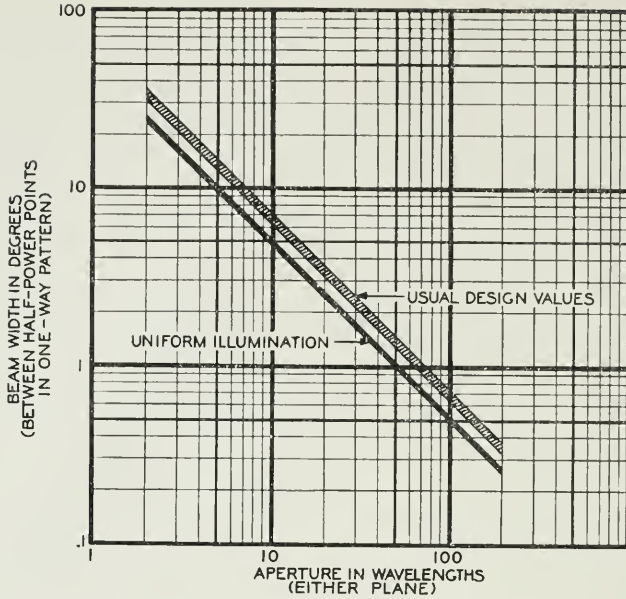
Fig. 1A—Antenna beam width vs. aperture
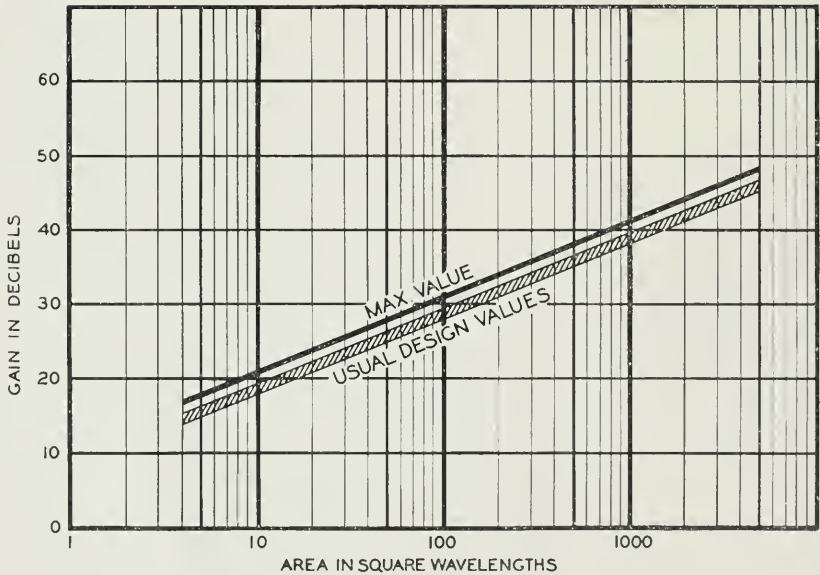
Fig. 1B—Antenna gain vs. aperture

and beam width obtained in this manner are shown by shaded area labelled *usual design values*. The need for higher frequency or shorter wavelength

was apparent to all of the early experimenters since physical limitations restricted the size of antenna which could be installed conveniently aboard ship. However, development effort along these lines had previously been hampered by lack of suitable vacuum tubes.

In spite of the vacuum tube difficulties the Laboratories work was started in the range from 500 to 700 mcs, a region several times that then in use at the Army and Navy laboratories. The best tubes available were those of the *doorknob* type which have been described in the literature by A. L. Samuel[2] and are illustrated in Fig. 2. The smallest of these was used in the receiver input circuits and two of the middle sized ones were used in the tramsmitter oscillator. These triodes operated at quite high frequencies by virtue of the very small spacing between their electrodes, a feature which made them fragile and demanded the development of plate modulation. Earlier radars had generally used grid keyed oscillators, i.e., the plate voltage was applied to the oscillator continuously together with a high grid bias voltage. The bias was removed momentarily by the keyer to emit a pulse. In order to obtain a useful pulse output from the doorknob oscillator tubes it was found essential to remove all stress from them except during the pulse. This was accomplished by using a direct coupled pulse amplifier or modulator, effectively in series with the oscillator and the power supply. Here again in 1938 no really suitable tubes were available for the modulator service since it also demanded a highly intermittent duty. However, since the modulator duty did not require the tubes to operate at very high frequency it was possible to use rugged high-voltage triodes which had been designed for continuous service, and to obtain the required pulse current capacity by paralleling a number of tubes. The earliest radar modulators used in the Laboratories employed a group of Eimac 100-TH tubes. Later, in the CXAS and Mark 1 Radars, six tubes similar to the W. E. 356A were used in parallel.

After a great deal of laboratory work an experimental equipment was assembled and demonstrated to the Army and Navy in July 1939. This early radar was notable in that it operated at what was then considered a very high frequency and also in that it employed a single antenna only about 6 ft. square. The transmitter and receiver were connected to the common antenna by a *duplexing technique* to be described later, which had been applied at lower frequencies by engineers at the Naval Research Laboratory. The results of these first field tests were encouraging and both the Army and the Navy ordered one prototype model equipment to be known as the CXAS. This radar was to operate at 500 or 700 mcs and was to incorporate a number of new features which were designed to make it

---

[2] *Proceedings of I. R. E.*, Vol. 25, page 1243, 1937—"Negative Grid Triode Oscillator and Amplifier for Ultra High Frequencies." Digest in Oct. 1937 *B. S. T. J.*
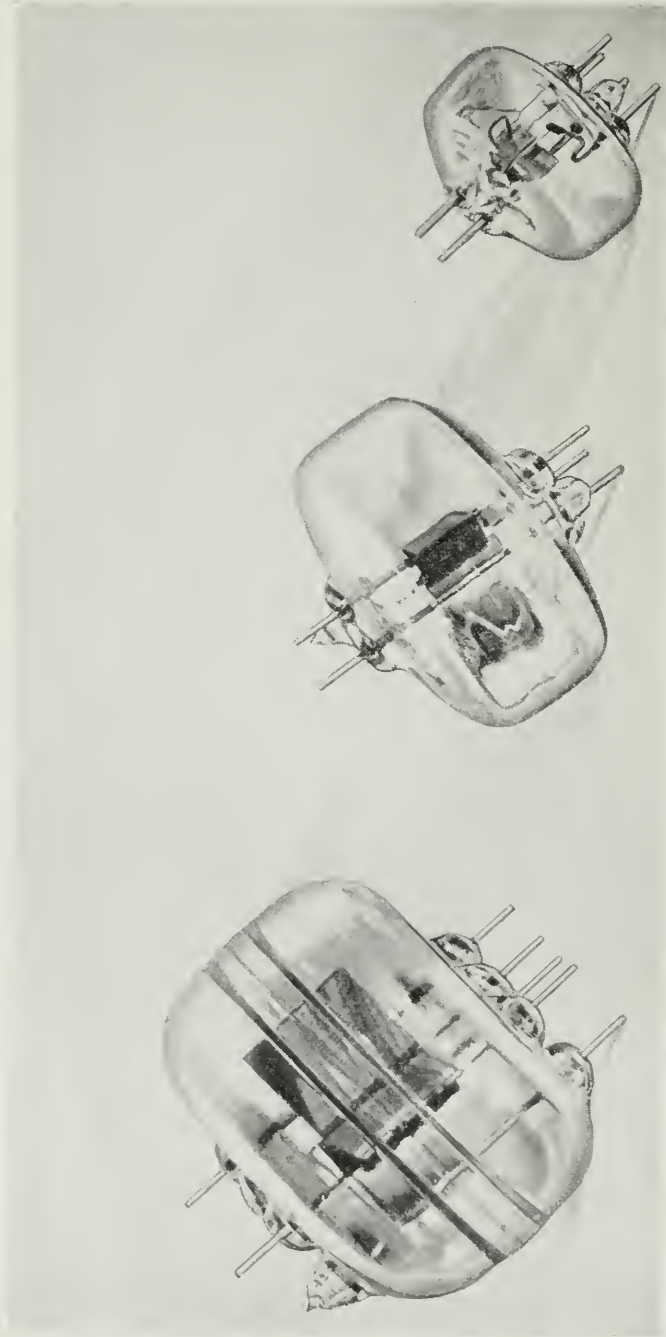
Fig. 2—Vacuum tubes for ultra-short waves

convenient to operate and to provide a range accuracy that would be useful in fire control. Since this early radar is of cosiderable historical importance it will be described in some detail.
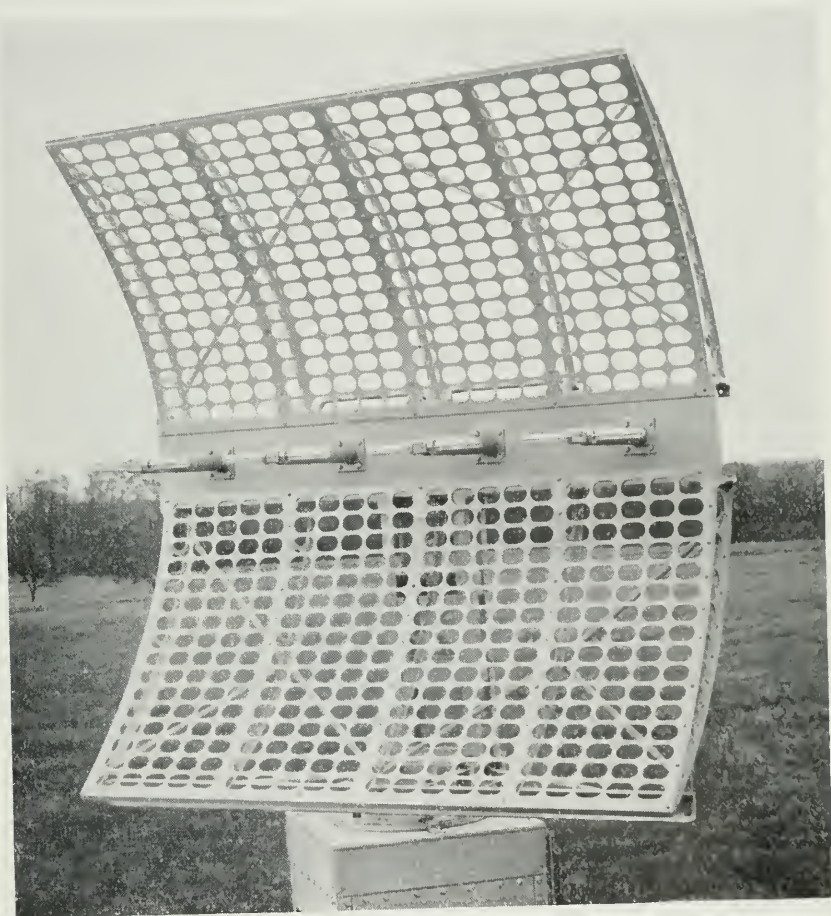
Fig. 3—CXAS—Antenna

The CXAS Radar

This equipment was divided into three major assemblies and the circuits were arranged so the three could be installed at some distance from each other. The antenna (see Fig. 3) consisted of a cylindrical parabolic reflector about 6 ft. square with an array of eight half-wavelength dipoles along the focal line. With shipboard use in mind the reflector was perforated to minimize wind resistance and the dipole and coaxial line feed

system was made weatherproof, which was accomplished by making the line system pressure-tight and filling it with dry gas. The gas-line system was extended to include the radiating elements by covering the latter with pyrex test tubes sealed to the support with a packing gland as shown in Fig. 4. A device was included in each dipole assembly for supplying the two half-wavelength radiating elements with balanced voltages from the unbalanced line, while a coaxile line harness including impedance matching
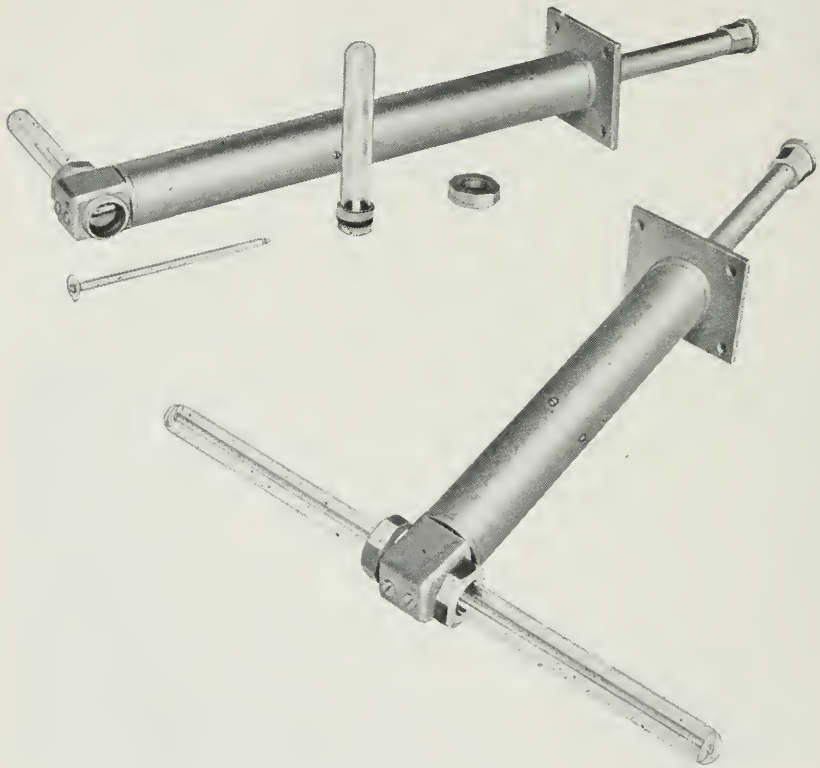
Fig. 4—CXAS—Dipoles

transformers was used to connect the several dipole assemblies and provide a matched load to the single transmitter-receiver line. A schematic diagram of this arrangement is shown in Fig. 5. The contemplated use of this radar was for surface targets or low-flying planes and rotation was provided only in azimuth. A gas-tight rotary joint was developed to carry the $\frac{7}{8}''$ coaxial line through the azimuth axis (Fig. 6).

The operator's cathode ray oscillograph indicator and all of the essential
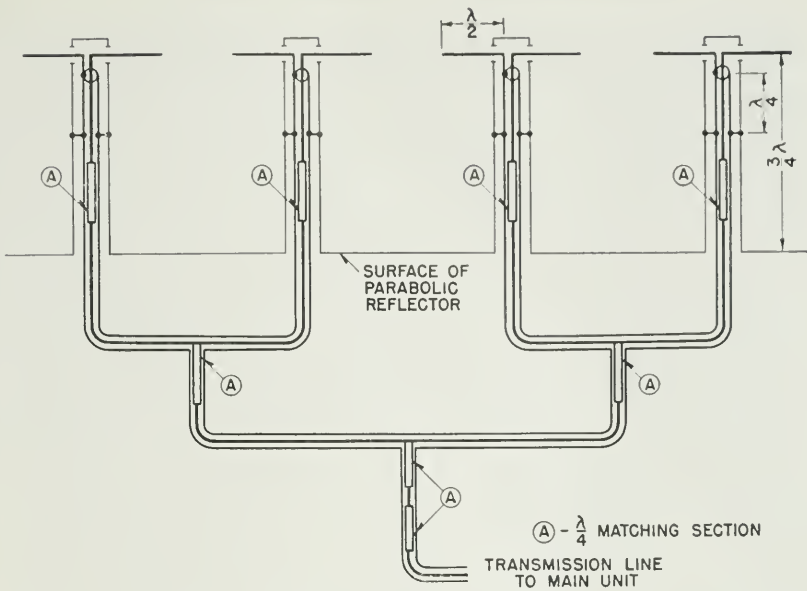
Fig. 5—CXAS—Antenna schematic
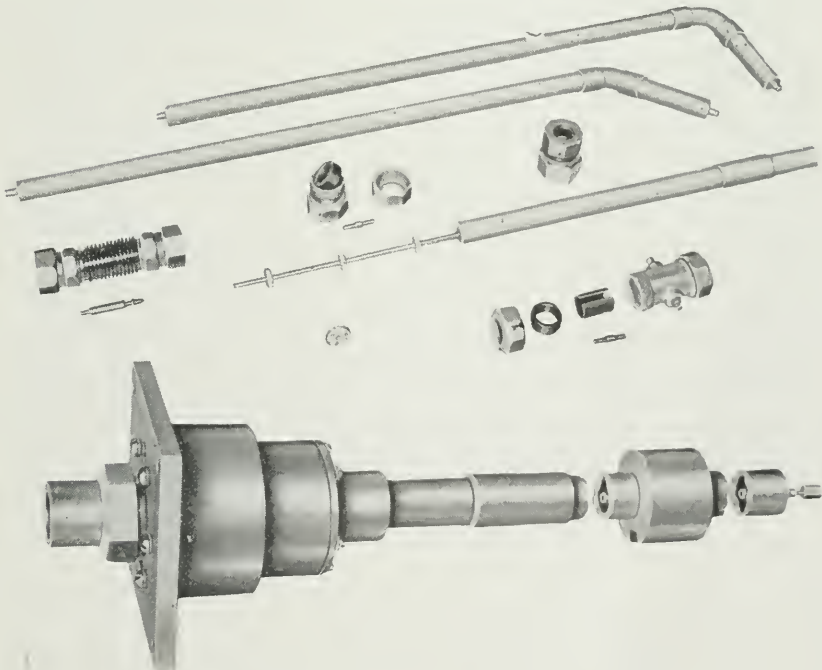
Fig. 6—CXAS—Rotary joint & transmission line fittings
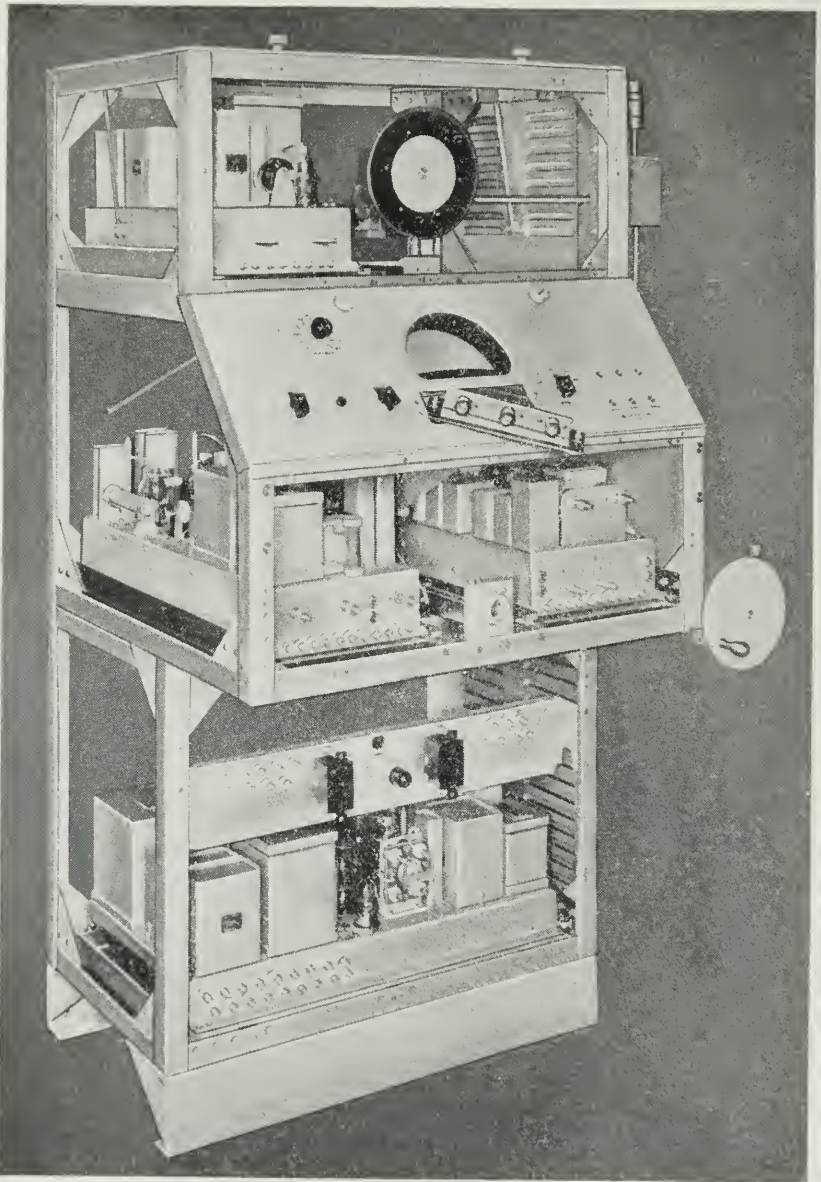
Fig. 7— CXAS—Indicator desk—covers removed

operating controls were combined into an assembly called the Indicator Desk, a photograph of which is shown in Fig. 7. This was intended for indoor mounting below the antenna in such a position that the azimuth

hand wheel on the desk could be connected to the antenna turntable by a shaft. The indicator employed a 7″ cathode ray tube and displayed the radar signals by what is now known as a Class A sweep with a full scale of 100,000 yards. A pioneering feature of this indicator was the provision of a series of electronic range marks to increase the accuracy with which target range could be read. Earlier indicators had used a ruled mask for the range scale and had suffered in accuracy due to parallax, sweep non-linearity, drift of sweep position, etc. The CXAS provided sharp pulses to mark the 10,000-yard intervals along the sweep line, and smaller pulses to mark the intervening 2,000-yard intervals. This system was free from the errors of the ruled mask and permitted range readings accurate to ±200 yards throughout the 100,000-yard scale. Provision was also made to expand any desired 20,000-yard segment of the scale to fill the entire tube screen so that signals could be examined more closely. The ranges corresponding to the 10,000-yard intervals were designated by illuminated
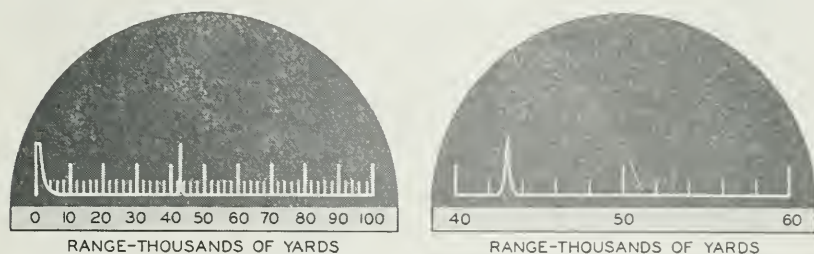
Fig. 8—CXAS—Range mark system

numerals located directly below the electronic scale. The presentation obtained with this arrangement is indicated in Fig. 8 which shows the electronic calibration marks, transmitted pulse, and an echo at 43,000 yards on both the full and expanded scales.

The third part of the CXAS equipment was an assembly known as the Transmitter-Receiver or Main Unit. It was designed to be unattended in normal operation and contained the Pulse Generator, Radio Receiver, Power Control Panel, Radio Transmitter, and H.V. Rectifier, which were all built as removable *drawer type* units. A side compartment in the Main Unit also housed the duplexing circuits, gas equipment for the transmission line, and some built-in test equipment, including a wavemeter and monitoring rectifier. The Main Unit and its sub-units are shown in Figs. 9 to 14, respectively. A single $\frac{7}{8}″$ coaxial transmission line provided connection from the Main Unit to the antenna.

In order to use a single antenna for both transmission and reception, means had to be provided to effectively disconnect the receiver during the
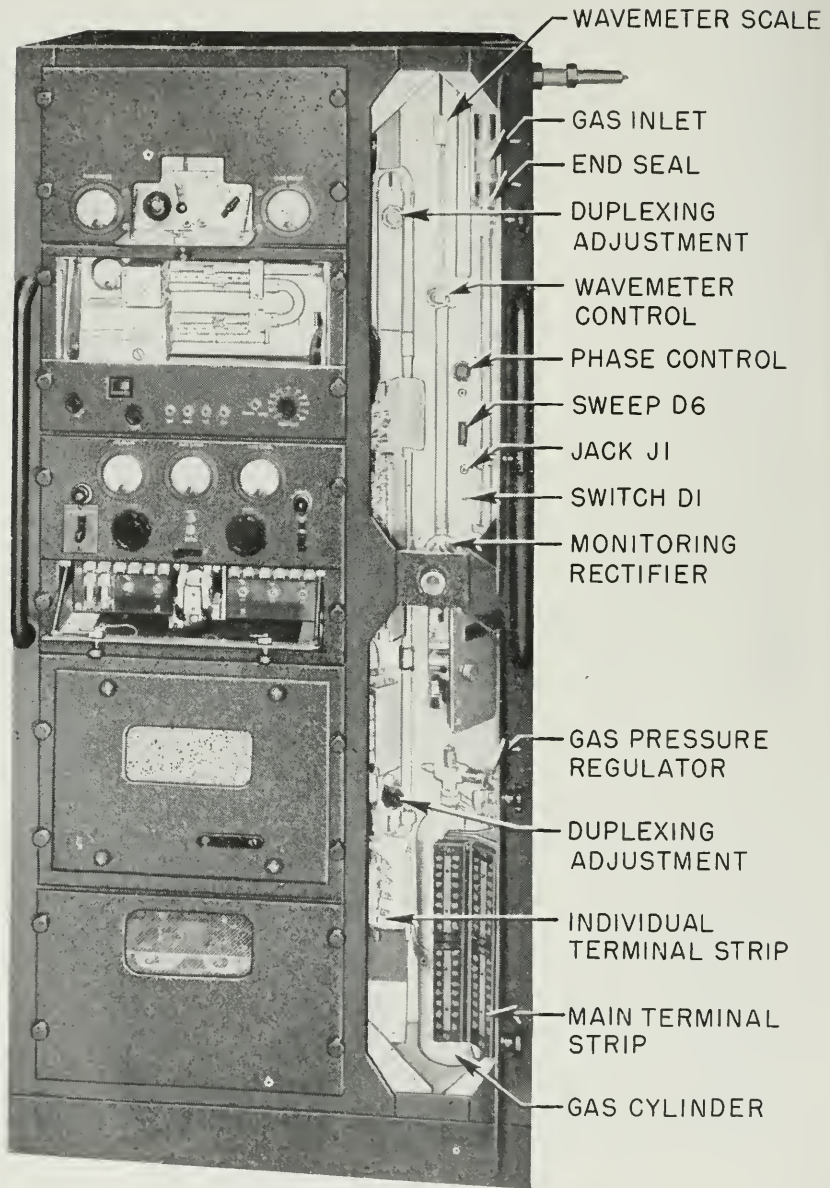
Fig. 9—CXAS—Main unit

transmitted pulse and to effectively disconnect the transmitter when the echo is received. If this were not done, a large part of the transmitted

energy would be dissipated in the receiver. Also, the minute received energy would be partially lost in the transmitter output circuit thus reducing the maximum range. Because of the extremely short time intervals between transmitted and received pulses, ordinary switching methods cannot be used. A *duplexing technique* mentioned earlier was therefore developed to provide this function. In the CXAS Radar this switching was obtained by connecting the transmitter and receiver to the antenna transmission line through adjustable lengths of coaxial line which were preset for a given operating frequency to be effectively an odd multiple of
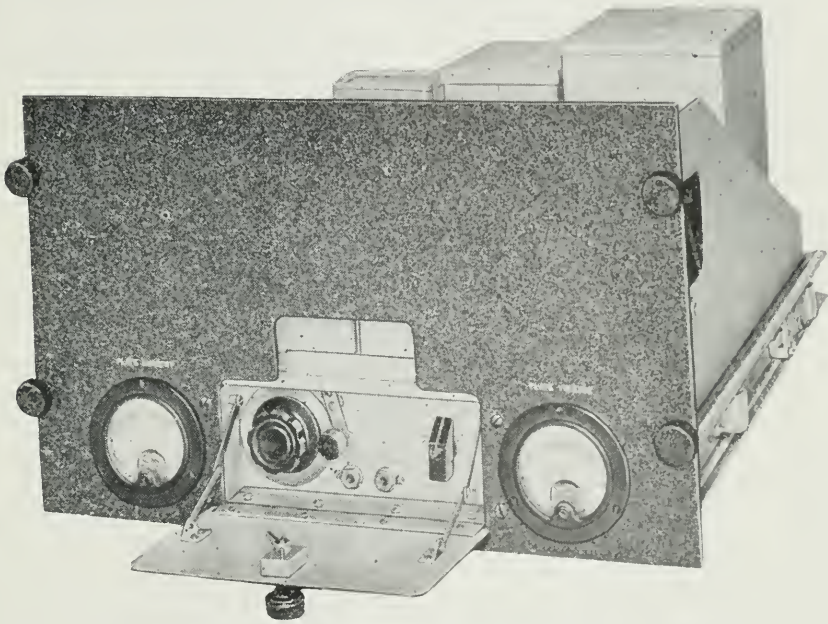
Fig. 10—CXAS—Modulation or pulse generator

one-quarter wavelength long. During the transmitted pulse, a small amount of the transmitted power overloaded the first tube in the receiver and provided a low impedance at that point. Due to the line length between receiver and junction point this low impedance is reflected as a high impedance at the junction point with the result that very little power is lost in the receiver line. At the end of the transmitted pulse the output impedance of the transmitter consists only of the small inductance of the output coupling loop and this impedance is reflected by proper choice of line length as a very high impedance at the junction joint with the receiver line. Thus, most of the received echo is directed to the receiver input
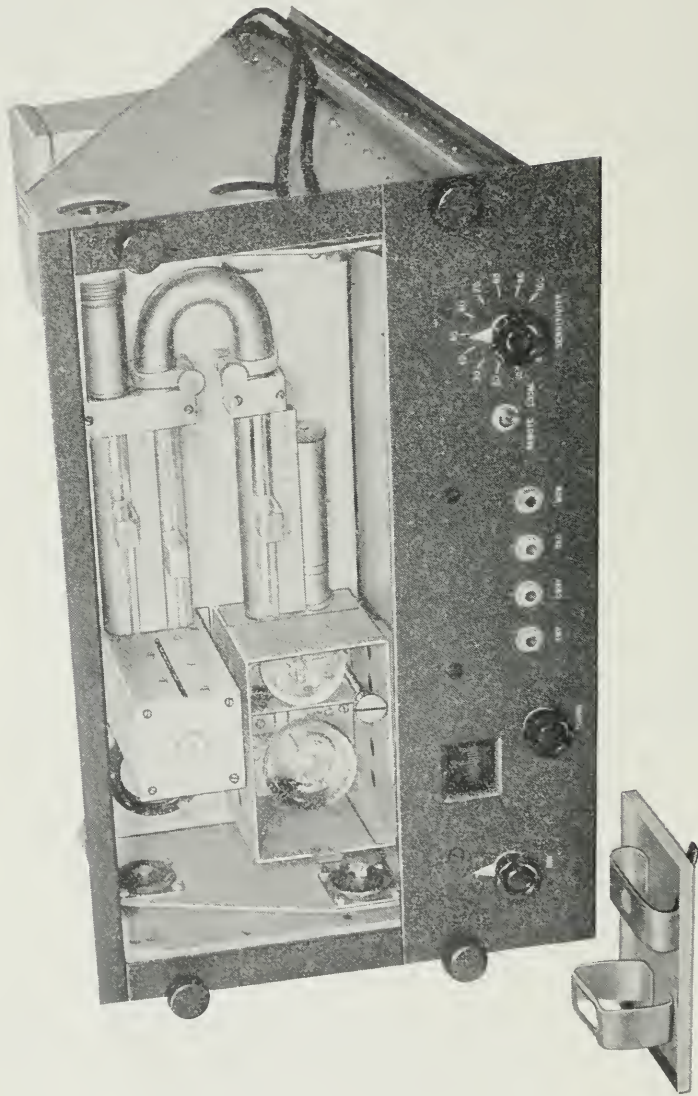
Fig. 11—CXAS—Receiver
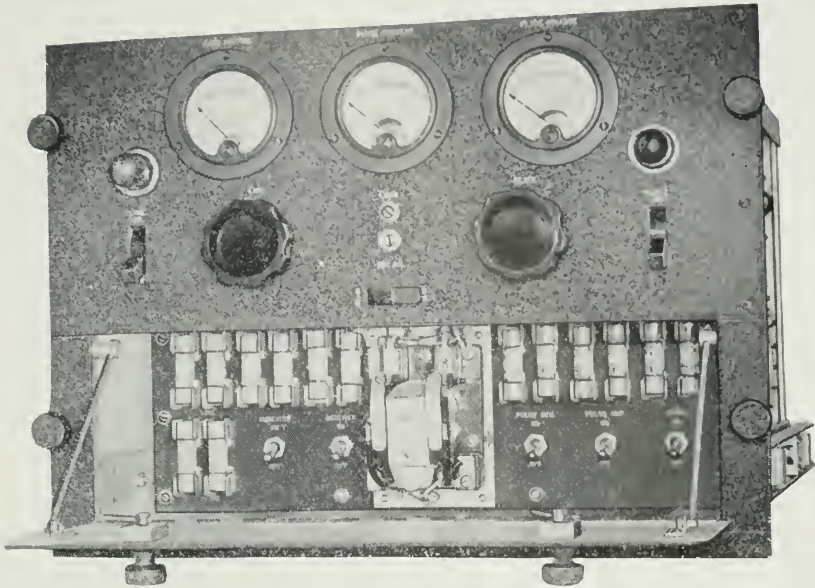
Fig. 12—CXAS—Power control panel
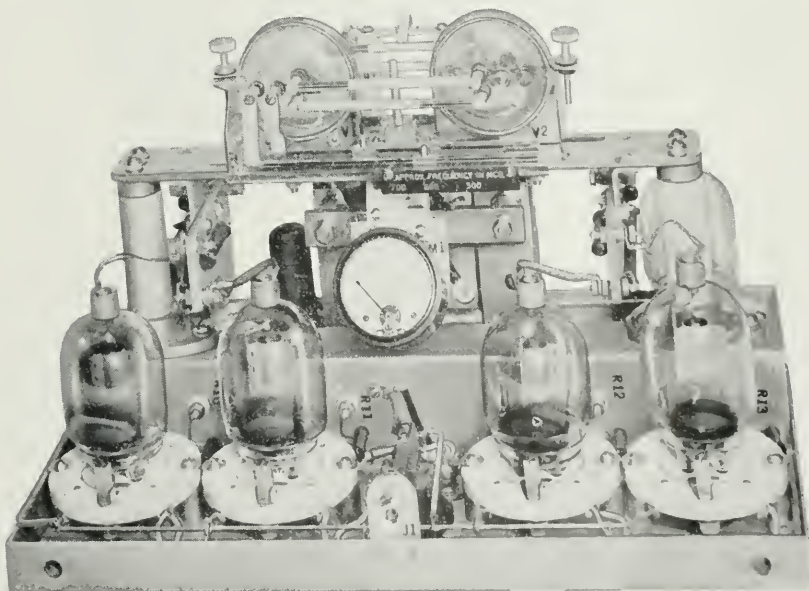
Fig. 13—CXAS—Transmitter

circuit. The adjustable duplexing transmission lines may be seen in the side compartment of the Main Unit on Fig. 9.

The equipment just described could be operated over a small frequency band of about 40 megacycles in the neighborhood of either 700 or 500 megacycles. The transmitter, receiver, and duplexing circuits were tunable over the entire range, but it was necessary to set up the antenna for one band or the other. This was accomplished by installing the proper one of the two sets of dipoles furnished, and installing or omitting a set of wedges
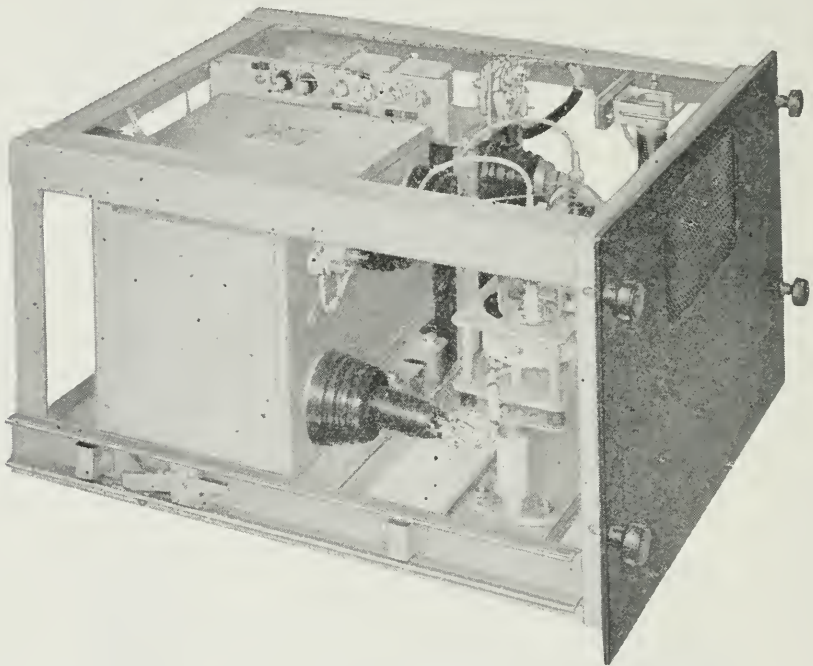
Fig. 14—CXAS—High voltage rectifier

which tilted the reflector wings to change the parabola focal length. This antenna set a precedent in design in that the dipoles and coaxial harness were designed for fairly broadband operation and were entirely free from field tuning adjustments which had been very troublesome in earlier equipment.

The CXAS Radar was demonstrated to the Navy in December 1940. After a few tests it was decided by the Navy to standardize on the 700-megacycle band. One of the principal reasons for this was that the tests had proved the superiority of shorter waves for surface target work; the

CXAS having regularly out-performed much higher powered equipment operating at 100 or 200 megacycles for this service. The reason for this can best be uuderstood by reference to Fig. 15 which illustrates what happens when a radio beam is directed horizontally over water. The beam breaks up into an interference pattern of several rays due to reflection from the surface; the position of the lowest ray depending only upon the height of the antenna measured in wavelengths above the water. Since the mount-

RADIATION PATTERN IN PRESENCE OF PERFECT REFLECTING SURFACE
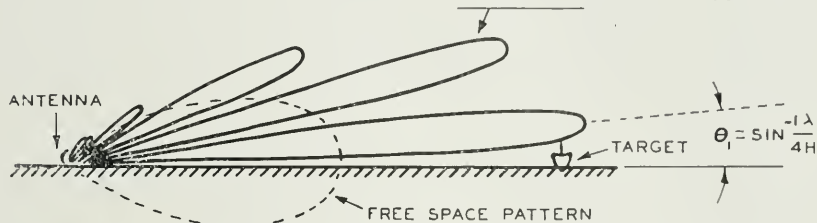
$$c = 2 \sin \left(2\pi \frac{H}{\lambda} \sin \theta \right) \times \left[\text{FREE SPACE ANTENNA PATTERN} \right]$$

WHERE   H = ANTENNA HEIGHT

λ = WAVE LENGTH IN SAME UNITS AS H

θ = ELEVATION ANGLE IN DEGREES

c = RELATIVE FIELD STRENGTH

Fig. 15—Effect of surface reflection on elevation beam

TABLE I

Operating Frequency	Tunable 680–720 mcs.
Antenna	Dipole array of 8 half-wave radiators, reflector 6' x 6', beam width 12 degrees, gain 22 db.
Transmitter Pulse Power	Approximately 2 kw.
Pulse Repetition Rate	1640 PPS
Pulse Duration	Variable in 5 steps from 1 to 5 microseconds.
Receiver-Superheterodyne	1 mc bandwidth, 30 mc IF frequency.
Receiver Noise Figure	Approximately 24 db.
Range Calibration	Electronic marks at 10,000 and 2,000 yard intervals.

ing height available aboard ship is fixed, the use of shorter wavelengths made it possible to keep the lowest ray more nearly horizontal where it could intercept a target's superstructure at greater distance.

The principal characteristics of the CXAS Radar as set up for operation at 700 megacycles are given in Table I.

This equipment gave useful results on surface targets at ranges of 10 miles or more (depending on the size of the target) and the range accuracy

of about ±200 yards was then considered very usable in surface target fire control. The target azimuth could also be determined to a precision of one or two degrees by rapidly swinging the antenna back and forth and observing the point which gave a maximum echo signal. This angular information was hardly good enough for fire control use. The equipment was also of some use against low flying aircraft as a means of getting better range data for fire control. Minor equipment difficulties were not entirely solved; in particular the doorknob triodes in the transmitter had a very short life under the high voltage pulse operating conditions. They had, of course, been designed originally for CW communication use and strenuous development effort to make them more suitable for the intermittent high power radar use had not been very successful.

The Mark 1 Radar

In spite of the obvious unsolved development problems the Navy immediately ordered 10 equipments, similar to the CXAS, for use in the Fleet. These were first called the FA Radio Ranging Equipment but the designation was later changed to Radar Mark 1. Several changes were made to better adapt the equipment for installation aboard ship, the principal one being a servo driven antenna pedestal of the amplidyne type which was furnished by the General Electric Company. The servo system eliminated the antenna drive shaft problem while retaining control from a handwheel on the control desk. The desk was also modified to provide dials reading both relative and true azimuth bearing, the latter being obtained by interconnection with the ships gyro compass system.

The first Mark 1 Radar was shipped by the Western Electric Company in June 1941 and installation on the USS Wichita was completed at the Brooklyn Navy Yard early in July 1941. This was the first fire control radar in our Fleet and the first of many thousands of radars of all types which the Western Electric Company was destined to build for the Navy in the following four years.

The Mark 2 Radar

While the ten Mark 1 radars were being built, development work was proceeding at top speed on major improvements designed to increase performance, eliminate operating troubles, and to make this new device fit better into the existing fire control situation aboard ship. The older optical devices were neatly integrated into a system, many features of which were automatic. For example, the gyro stabilized telescopes and optical range finder were assembled into a compact rotating armored box called a *director*, located high on the ship. Target data from the director was sent

automatically by *synchro* data transmitters to the computer below decks, which solved the fire control problem and likewise transmitted automatically the correct information to the guns. For the new radar target locating device to fit into the existing system it was necessary to make its angle finding function operate more in the manner of the telescopes. Not only was it desired to determine target angles more accurately but it was necessary to track target position continuously and smoothly. Finally, to take care of the anticipated need for rapidly changing back and forth during an engagement from optical to radar data it became apparent that the same operators should handle both jobs. Thus it was decided that the system should provide the existing operators with oscilloscopes to supplement their telescopes, and to arrange them so either could be used as desired. Further to coordinate the data it became obvious that the radar antenna should be connected with the optics in such a way that the two were always pointed in the same direction. This would make it possible to leave the existing data transmission system alone and would avoid any break in data when changing from optics to radar or vice versa. For example, if a visible target disappeared behind a fog bank the telescope operator would simply move his head to look at his oscilloscope and data would continue to flow smoothly to the computer and to the guns.

Thus the engineers of the Navy decided the new radar device could be fitted into the existing fire control system. Any other decision would likely have required modification of many parts of the system, and would have delayed the extensive use of fire control radar by a matter of years. The Bell Telephone Laboratories were accordingly asked to modify and improve the radar design to make possible the coordination of optics and radar as just discussed. The new radar was to be called Mark 2 and was to be similar to the Mark 1 but modified to provide continuous tracking in azimuth with an accuracy of ± 15 minutes of arc, and continuous tracking in range with an accuracy of ± 50 yards. Further, the operator's oscilloscopes and controls were to be put into small units that could be mounted alongside of the telescopes in the director, and the antenna was to mount on the director. These requirements demanded some important forward steps in radar development which will be described in some detail. Before Radar Mark 2 got into production a much higher powered transmitter was developed and with this change the equipment was re-named Radar Mark 3.

THE MARK 3 RADAR

The general arrangement of apparatus for this radar differed from the Mark 1 principally in the indicators, which were designed to mount in the

already crowded gun director. These indicators are shown in Figs. 16, 17
and 19. Fig. 16 shows the range operator's oscilloscope, called the Control
and Indicator, which was located near the optical range finder. Adjacent
to this unit was mounted the range unit, shown in Fig. 17 by means of
which the operator could select the target to be followed and continuously
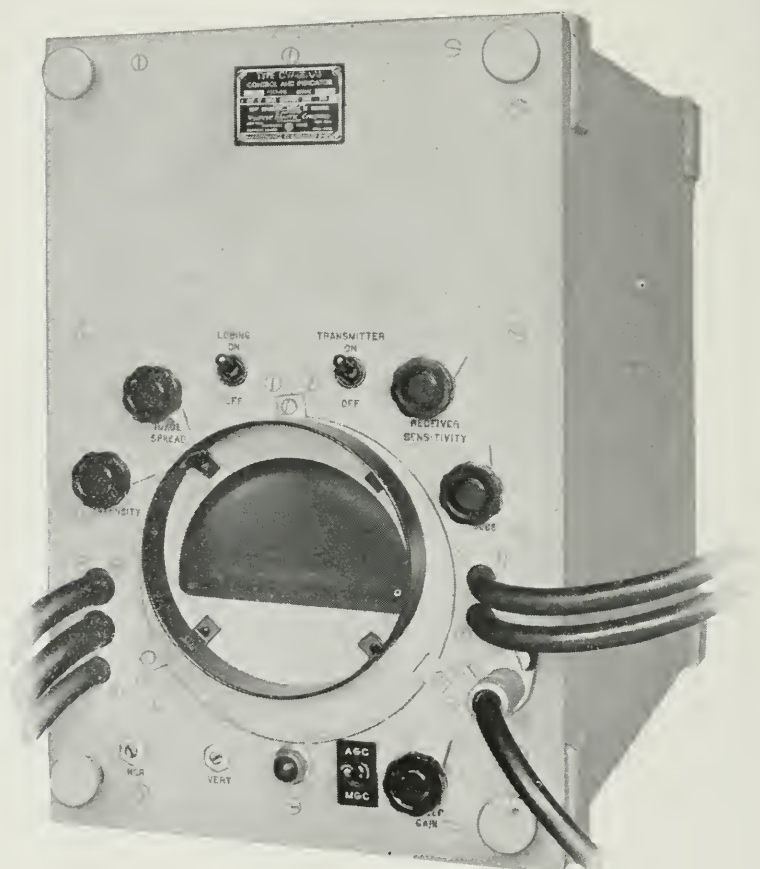
Fig. 16—Control & indicator—Radars Mark 2, 3 & 4

maintain accurate range readings. A typical installation of these two
units is shown in Fig. 18. The third unit, shown in Fig. 19, is called a
Train or Elevation Indicator and, in Radar Mark 3 (which was for sur-
face fire only) this indicator was mounted adjacent to the Train (azimuth)
Operator's telescope.

In addition to the Train Indicator, the azimuth operator was provided
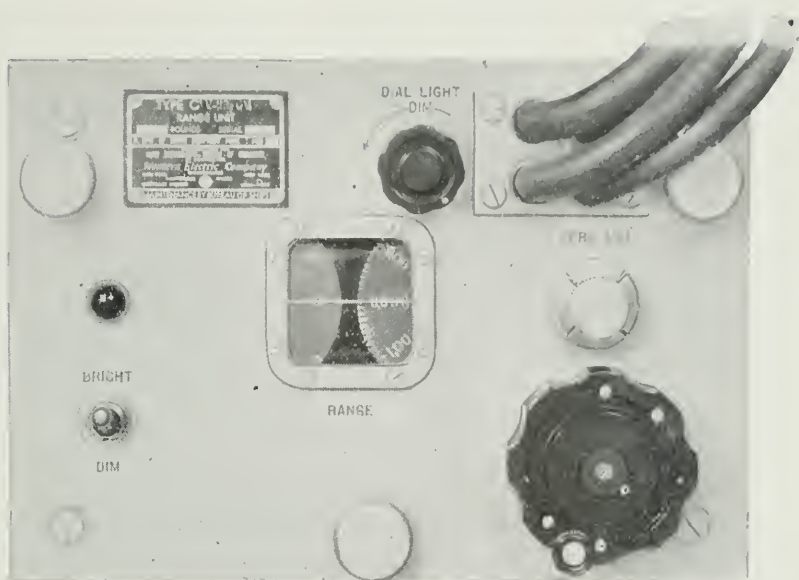
Fig. 17—Range unit—Radars Mark 2, 3 & 4
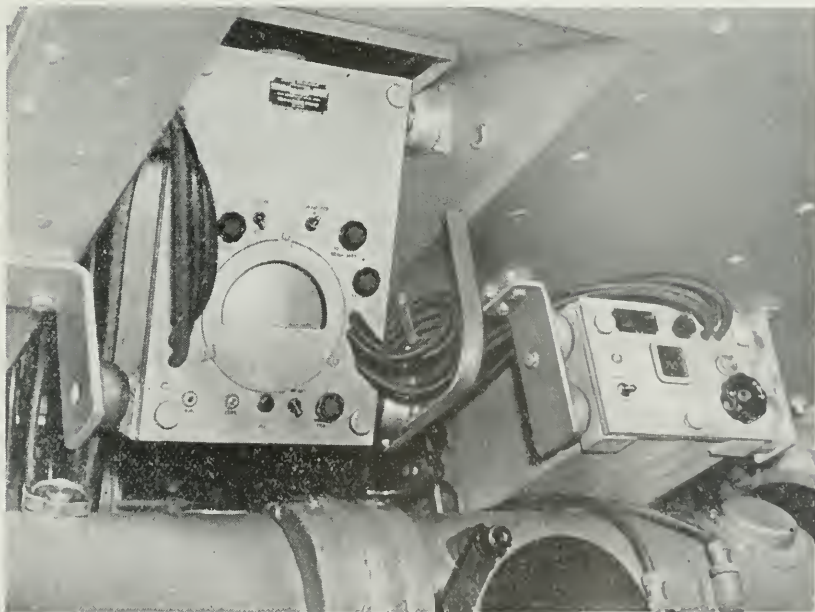
Fig. 18—Mark 3 Radar—Range operator's position on Cruiser Honolulu (Navy Photo 153-6-42)

with a Train Meter of the zero center type which indicated the direction of deviation from true target position.   One of these meters of early design can be seen in Fig. 38.   Two meters of later design are shown in Fig. 39 mounted immediately below optical telescopes.   The pulse generator, receiver, transmitter, rectifiers, etc., were located below decks in the Trans-
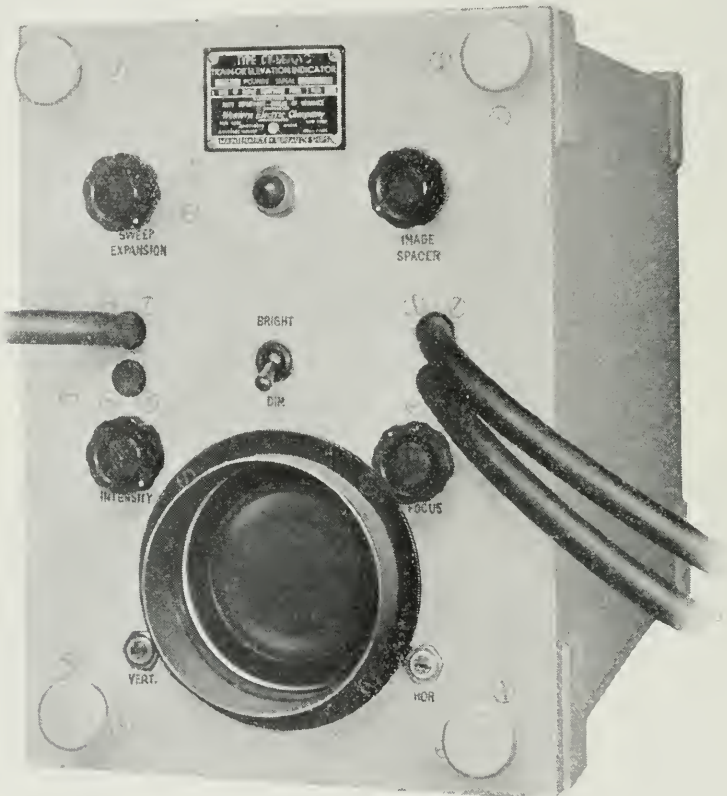
Fig. 19—Train or elevation indicator—Radars Mark 2, 3 & 4

mitter-Receiver, or Main Unit which was very similar in appearance to the Main Unit of Radar Mark 1 shown in Fig. 9.

Two types of antennas were provided for this radar: a 6 ft. by 6 ft. parabolic array similar to the Mark 1 antenna, and a 3 ft. by 12 ft. parabolic array.   Either one or the other of these antennas was mounted on top of the gun director and rotated with it in azimuth.   Both were provided with azimuth lobe switching to be described later.   Because of the relatively narrow elevation beam of the 6 ft. by 6 ft. array, this antenna required gyro

stabilization in elevation to take care of pitch and roll of the ship. Such stabilization was not required with the broad elevation beam obtained with the 3 ft. by 12 ft. antenna; and in addition, this wider antenna provided more accurate tracking due to the narrower antenna beam in azimuth. Installations of these antennas aboard ship are shown in Figs. 20 and 21 and 22, 23 and 24.
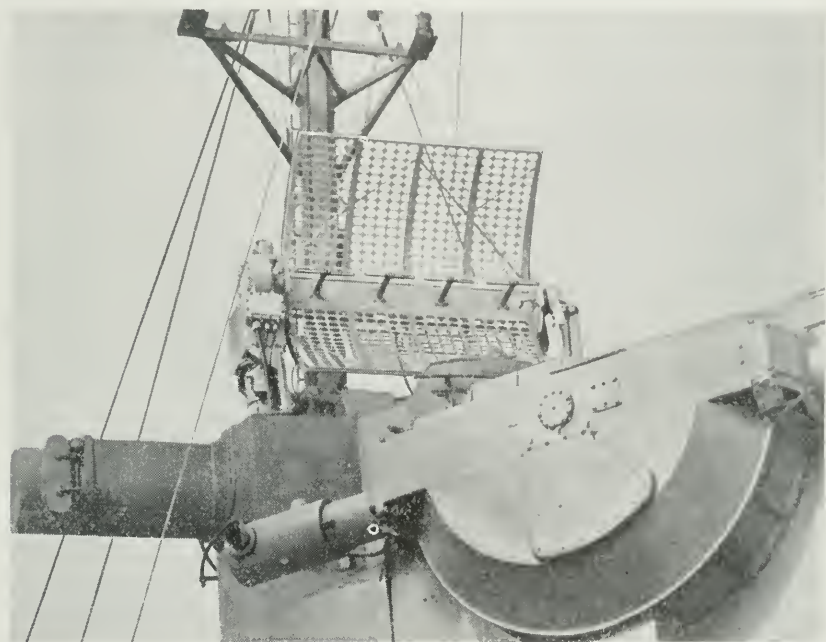
Fig. 20—Radar Mark 3 antenna (6′ x 6′) on Cruiser Honolulu (Navy Photo 144-6-42)

Antenna Lobe Switching

The problem of measuring angles accurately with a relatively broad radio beam has been faced many times in the radio direction finding art. The most successful attack has made use of the fact that while the nose of a radio antenna beam is blunt, the sides of the beam are relatively steep; i.e., while the rate of change of signal amplitude with angle is very low near the nose of the beam it becomes substantial down on the side of the beam. A very well known application of this principle is the airway radio range wherein two very broad overlapping beams define a narrow path by utilizing the points where the two overlap with equal intensity. A somewhat similar scheme in which the antenna beam is switched rapidly between two positions has been applied in radar, and in an early form was first used

in this country by the Signal Corps in the work described by General Colton, to which reference has been made.

The use of two antenna beam (or lobe) positions to obtain more accurate radar angle data is referred to as *lobe switching* and the operating principle is illustrated in Fig. 25. The antenna beam is shown in two positions: position 1 being directed to the right, and position 2 to the left of the mechanical axis of the antenna. The antenna beam is caused to switch rapidly between these two positions, and simultaneous with this switching a small horizontal displacement of the indicator Class A sweep is introduced. In
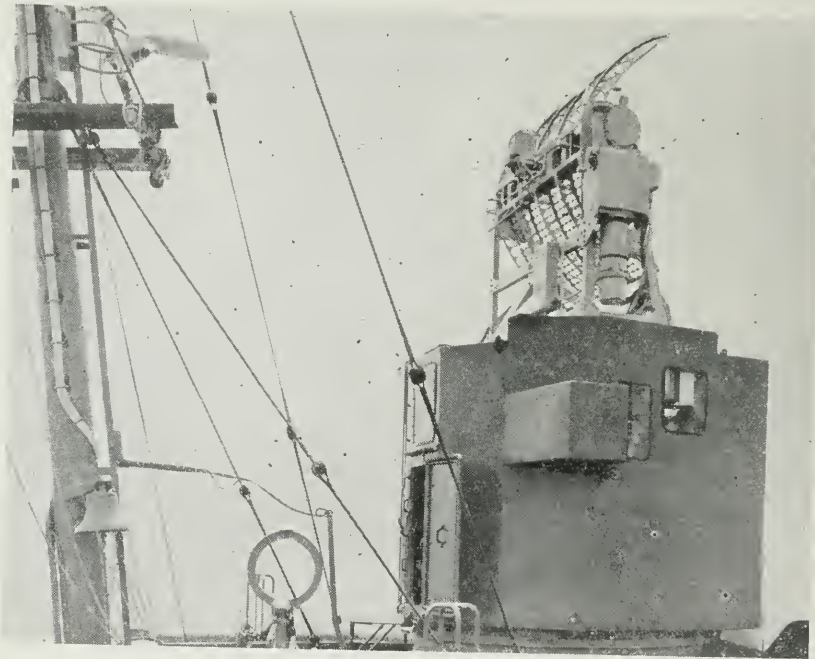
Fig. 21—Radar Mark 3 antenna on Destroyer Porter (Navy Photo 2711-42)

this manner the signals received in the two beam positions may be viewed separately. The speed of switching is made sufficiently high to minimize flicker and the effect of fading signals. It will be noted from this diagram that the signal strength received from target A is the same for both beam positions thereby producing equal "pip" heights on the indicator screen. However, for target B the signal amplitude is greater in position 1 than in position 2 and the "pip" amplitudes on the indicator differ correspondingly. If the operator wishes to track target B it is only necessary for him to rotate the antenna until the two "pips" are of equal amplitude. Smooth
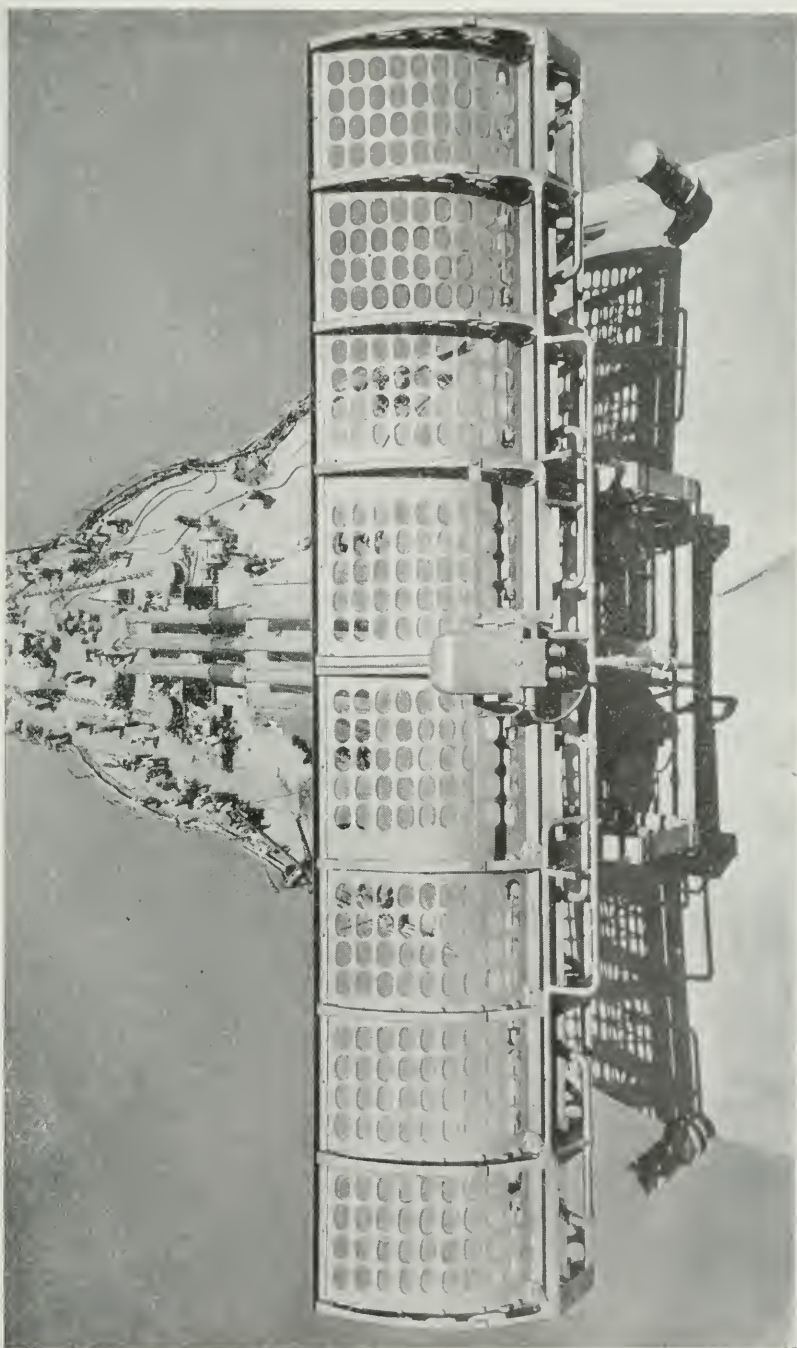
Fig. 22—Radar Mark 3 antenna (3′ x 12′) on Battleship Pennsylvania (Navy Photo 4273-42)
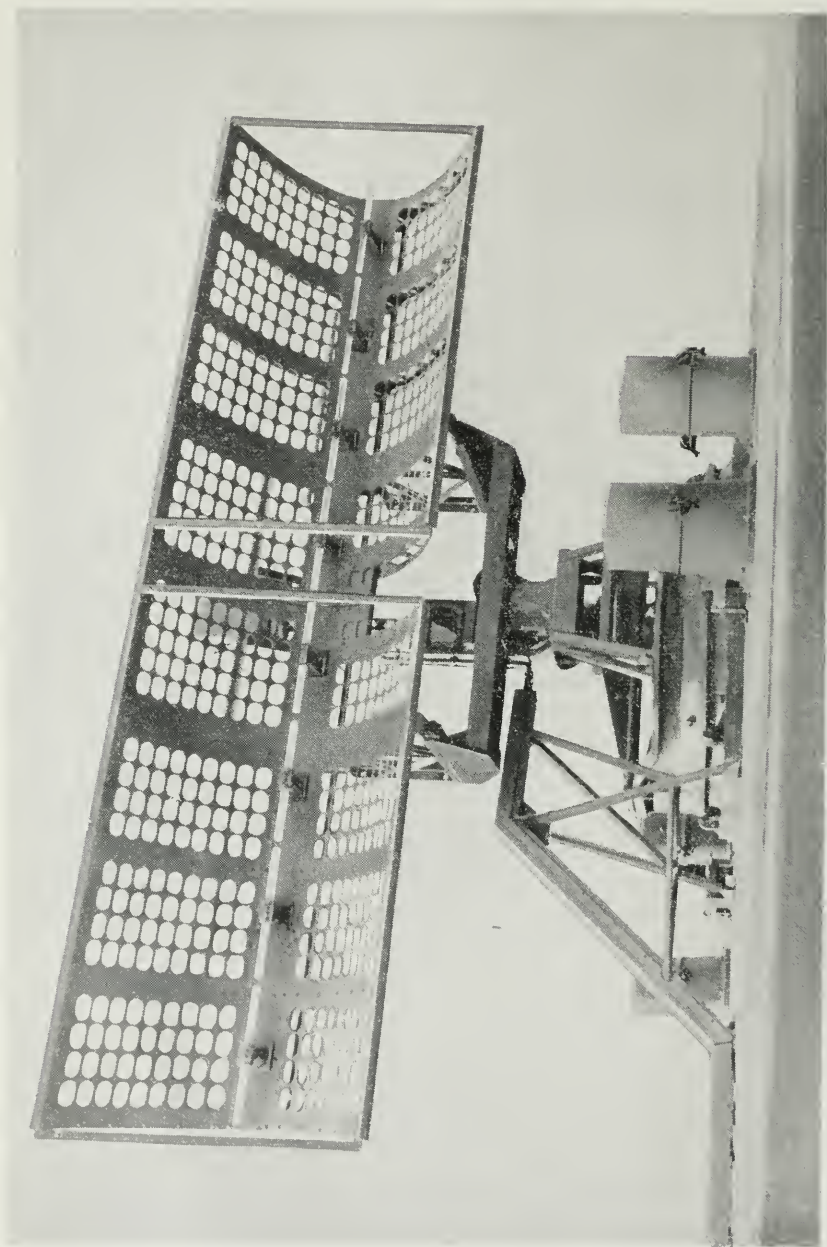
Fig. 23—Radar Mark 3 antenna (3' x 12') on Battleship New Jersey. (Navy Photo 181812)

flow of azimuth data will be obtained if the operator continuously maintains equal amplitude of the two "pips".

In the Signal Corps equipment to which reference has been made, separate antennas were used for transmission and reception with lobe switching
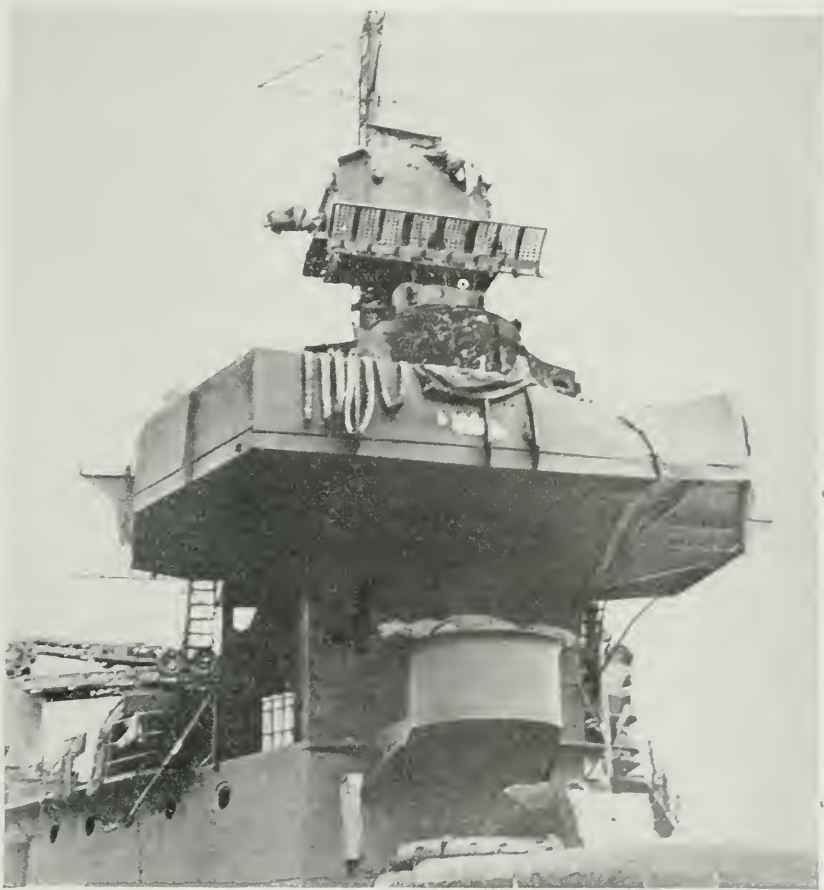
Fig. 24—Radar Mark 3 antenna (3' x 12') on Cruiser San Francisco after Pacific battle
(Navy Photo 34133)

applied only to the receiving antenna. Space limitations aboard ship made it mandatory to accomplish all functions using a single antenna. This required the development of a lobe switching device capable of withstanding the high peak power during the transmitted pulse; a problem which had not been faced in the Signal Corps equipment. It was further desired to provide a weatherproof lobe switching device, free from radio
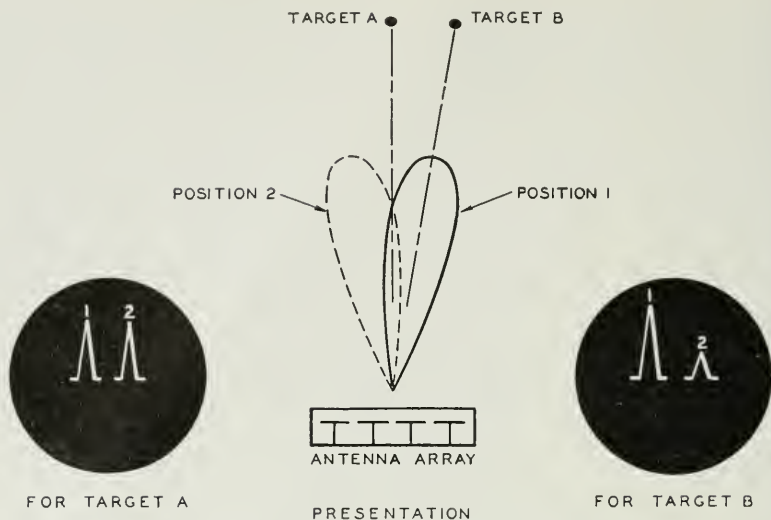
Fig. 25—Principle of lobe switching
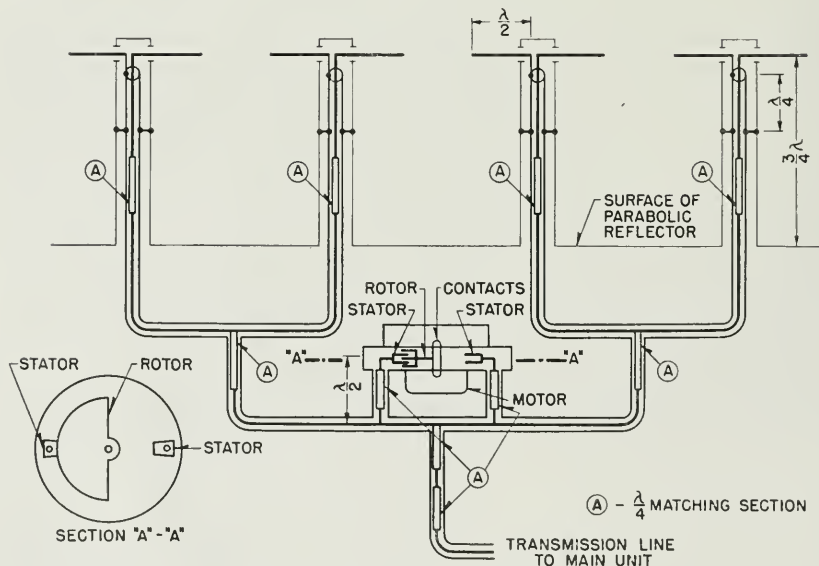
Fig. 26—Mark 2 & 3—Antenna schematic

frequency adjustments, in order to simplify operation and maintenance. The manner in which these objectives were met is described below.

To obtain lobe switching of the antenna beam, use was made of the fact

that the beam position depends upon the relative phase of the excitation applied to the radiating elements of the array. If all elements are excited in phase, as in Radar Mark 1, the beam will be normal to the line of the array, while gradually increasing phase difference across the array will result in displacement of the beam. For small angles of beam shift, entirely satisfactory results may be obtained by shifting the phase of excitation applied to one-half of the array with respect to the other, and this expedient results in a much simpler phase shifting mechanism than would be required to obtain uniform phase change. This system was used in Radar Mark 3 and its application is illustrated schematically in Fig. 26. It will be seen that this array is identical to that used for Radar Mark 1 except for the central section of transmission line in which a lobe switching unit has been added. In this unit the phase of excitation to one-half of the array is retarded with respect to the other half by connecting a capacitive reactance alternately across one feed line or the other to obtain the two beam positions. Switching is accomplished by the use of a motor driven rotary capacitor shown in Section A-A. The rotor is a semicircular aluminum casting which is maintained at substantially ground potential by very close spacing to the grounded metal housing. The two stators are small metal plates which interleave with the rotor during approximately one-half revolution and are connected through half-wavelength coaxial lines to the antenna transmission lines. The purpose of the half-wavelength stub lines is to avoid physical limitations which would otherwise be encountered in connecting the rotary capacitor to the lines. Allowance is made in these stubs for end-loading caused by stray capacitance of the stator plates and supporting insulators. It will be seen that during nearly one-half revolution of the rotor one of the stators is engaged to shift the antenna beam in one direction while during the other half revolution the other stator is engaged to produce the other lobe position. The switching occurs during the small interval in which both stators are engaged by the rotor. Signals received during this interval are blanked out in the indicator. The rotor of the lobe switcher is driven at about 30 RPS by an induction motor mounted within a weatherproof housing. The motor shaft also carries cam operated contacts to produce image spacing on the indicators, control signals for the Train Meter, and blanking during the lobe switch interval. The entire unit is gas tight and is filled with dry gas through the transmission line.

The value of the lobe switching capacitor and its position along the feed line must satisfy two conditions: first, the phase shift must be such that the antenna beam will be displaced by the desired amount; and second, the impedance at the feed point must be such that equal division of power will be obtained in the two halves of the array. In the first Radar Mark 3 antenna (6 ft. by 6 ft. parabolic array) a beam displacement of about 3.0

degrees was chosen as a suitable compromise between target angle sensitivity (steepness of beam) and reduction of signal amplitude "on target". This displacement required a phase shift of approximately 53 degrees between the two halves of the array.   From transmission line theory it can be shown that this phase difference will be obtained with a capacitive reactance equal to the characteristic impedance of the feed line when connected at a point 0.176 wavelength from the feed point.   It can also be shown that this condition satisfies the requirement for equal power division to the two halves of the array.   A capacitor of the required value (about 3 micromicrofarads) can readily be built to withstand the peak transmitted power by proper condenser plate separation.   A frequency variation of about 40 megacycles can be tolerated without materially affecting the antenna performance.

TABLE II.—*Antenna Characteristics*

	Radar Mark 3		Radar Mark 4
Dimensions	3' x 12'	6' x 6'	6' x 7'
Apperture in Wavelengths			
Azimuth	8.5	4.25	4.25
Elevation	2.1	4.25	4.95
Beam Width in Degrees (between half power			
points in one way pattern)			
Azimuth	6	12	12
Elevation	30	14	12
Antenna Gain in db	22.0	22.0	22.5
Beam Shift in Degrees			
Azimuth	$\pm 1.5°$	$\pm 3.0°$	$\pm 3.0°$
Elevation	—	—	$\pm 3.0°$

A lobe switching unit similar to that described above was also applied to the 3 ft. by 12 ft. antenna.   Pertinent information regarding beam widths and lobing angles for both antennas (together with information on the antenna for Radar Mark 4 to be described later) is given in Table II.

The effective beam widths as used in these radars were somewhat narrower than the values given above due to the square law characteristic of the second detector in the receiver, and the deflection sensitivity was such that the specified tracking accuracy of ± 15 minutes of arc could readily be achieved.   The "on target" position or axis of the antenna (lobe crossover) was carefully aligned with the optical telescopes at the time of installation so that either optics or radar angles could be used.   The symmetrical design of the antenna made this alignment substantially independent of small changes in operating frequency.

To minimize target confusion the signals presented on the Train or Elevation Indicator (azimuth operator's oscilloscope) consisted only of

those received from the target being tracked by the range operator, all others being blanked out in the indicator circuits.

Accurate Range Measurement

The second major problem which required solution to adapt radar to the fire control problem was the provision of means for accurate and continuous range tracking. It was obvious that what was required was some sort of electronic range mark on the indicator sweeps, the position of which could be varied by a rotary device whose motion could be used to transmit range information to a remote point over a synchro system. The range mark could then be aligned with the target "pip" on the oscilloscope. For accurate data transmission it was necessary to obtain a linear relationship between angular rotation of the range handwheel and corresponding range to the marker on the radar indicator screen.

One method which was first employed by the Signal Corps made use of the fact that the transmitted pulses were generated at a periodic rate from a sine wave oscillator of fixed frequency; the pulse being produced at a fixed point in each cycle. By transmitting this same sine wave through a linear phase shifter a new pulse could be generated whose position in time, relative to the transmitted pulse, could be varied by rotation of the phase shifter. In the Signal Corps equipment a special goniometer was used to produce the phase shift and the accuracy obtained was considered adequate for the intended purpose. However, non-linearity of the phase shifting device, though small, was much greater than could be tolerated in the Navy fire control system. A study indicated that large scale manufacture of special phase shifters, hand adjusted to meet the stringent accuracy requirements was out of the question. It was therefore decided that a two speed system be used, in which the phase shifter errors would be divided by the gear ratio to the high-speed unit in much the same way that accurate synchro information is transmitted by a "coarse" and a "fine" synchro. The manner in which this was worked out by Bell Telephone Laboratories and applied to Radars Mark 3 and 4 is described below.

The method of range measurement can perhaps best be understood by first examining the method of presentation used on the cathode ray tube indicator for the range operator. This presentation is shown in Fig. 27 in which it will be noted that a Class A sweep is used to display the transmitted pulse and received echoes. This horizontal sweep, however, differs from the simple sweep of earlier radars in several respects. First, the central portion of the sweep is expanded to permit more accurate viewing of signals appearing within this region; second, a downward deflection called the range "notch" is produced in the approximate center of the expanded section; and third, the circuits are so arranged that the notch

remains centered as the range unit phase shifters are rotated thus causing all of the signals (rather than the notch) to move across the screen. Range measurement is made by rotating the range unit handcrank to place the desired signal in the center of the range notch on the indicator. This type of presentation has several advantages. It permits the full 100,000-yard range to be viewed at all times so that new targets may be immediately detected, and permits accurate viewing of the desired target in the expanded
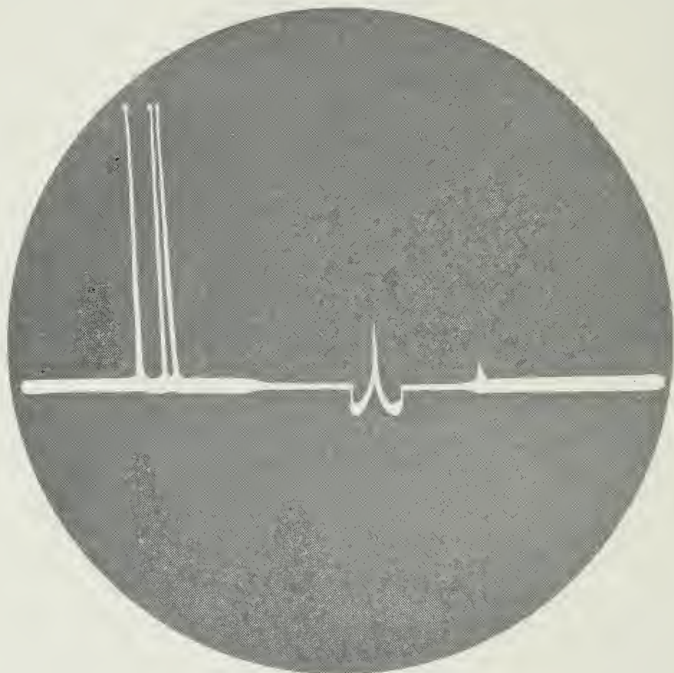
Fig. 27—Mark 3 & 4—Range presentation

center of the sweep where best focus is obtained. For smooth range tracking it is only necessary for the operator to rotate the range unit hand crank to keep the desired signal centered in the range notch.

A block diagram of the range measuring system, together with the circuits used to obtain the cathode ray indicator presentation described above, is shown in Fig. 28. A base or reference oscillator generates a sine wave of 1.639 kc, one cycle of which corresponds to a radar range of 100,000 yards. This wave, after amplification, is applied to a non-linear coil pulse generator[3] which generates short pulses (one positive and one negative pulse

[3] "Magnetic Generation of a group of Harmonics," E. Petersen, J. M. Manley, L. R. Wrathall—August 1937, *B. S. T. J.*, October 1937.
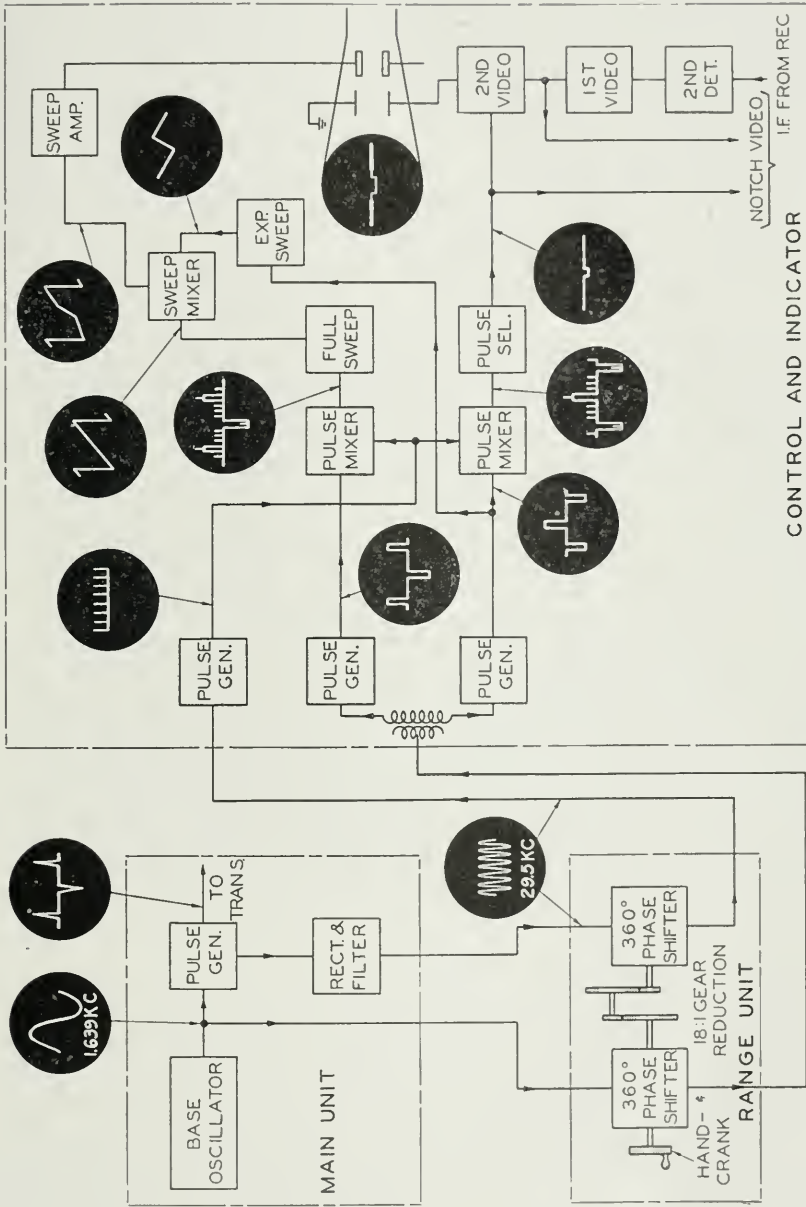
Fig. 28—Mark 3 & 4—Range measuring system

per cycle); the positive pulses being used for keying the transmitter. These pulses are rich in odd harmonics of the base oscillator frequency. By rectifying these pulses to reverse the negative pulses, even harmonics of the base frequency are obtained and the 18th harmonic (29.5 kc) is selected by means of a filter. This harmonic frequency and the original base frequency are applied to two phase shifters whose shafts are geared together in the ratio of 18 to 1. Since one revolution of the one speed phase shifter corresponds to 100,000 yards, one revolution of the 18-speed unit corresponds to only 5550 yards with the result that range errors caused by non-linearity of this phase shifter are reduced by a factor of 18. The phase shifters employed are similar to those designed by Bell Telephone Laboratories for use in a phase measuring bridge[4] and are linear to within ±1.5 degrees or about 0.4 per cent. The possible range error introduced by imperfections in the 18-speed phase shifter was therefore only 23 yards, well within the design requirements. It remains to be shown how this accurate range information was applied to the indicator.

The output of the 18-speed phase shifter in the range unit is connected to the Control and Indicator where the phase shifted sine wave is used to generate short, rectangular pulses of about 600 yards duration. One pulse is produced for each cycle of the 29.5 kc wave so that 18 of them occur during the 100,000-yard sweep interval. It is desired that only one of these pulses appear as a range notch on the indicator screen and this pulse is selected from the others by a pedestal pulse generated from the output of the one speed phase shifter. It will be noted that as the phase shifters are rotated by means of the range unit hand crank, the desired pulse from the 18-speed phase shifter will remain substantially centered on the one-speed pedestal pulse. After further shaping, the selected pulse is mixed with the received signals in the second video amplifier and is then applied to the vertical plates of the cathode ray indicator to form the "range notch". The range notch is also transmitted to the Train Indicator and Train Meter where it is used to prevent any signal from affecting those instruments except the one being tracked by the range operator.

Since it is desired to have the range notch appear in the center of the 100,000-yard sweep on the indicator, the sweep trigger pulse must occur 50,000 yards in advance of the notch. This trigger is obtained by selection of another pulse from the accurate phase shifter, this time using a one-speed pedestal produced by an input of reversed phase. The pulse thus selected is used as a trigger for starting a saw-tooth sweep wave with a duration corresponding to 100,000 yards radar range. Expansion of the center portion of this sweep is obtained by adding to this wave a second

[4] L. A. Meacham, U. S. Patent 2004613.

saw-tooth wave having maximum rate of change in the center of the sweep; the latter being derived from the range notch selection pedestal. The combined sweep is then applied to the horizontal plates of the cathode ray tube. The return trace is blanked by applying to the control grid of the cathode ray tube a voltage obtained by differentiating the sweep waveform.

Transmitter

As mentioned earlier the transmitter oscillator tube problem was one of the major obstacles in the march of radar development to higher frequencies. Intense development effort on many possible types of tubes was underway in several laboratories in this country and abroad during 1939
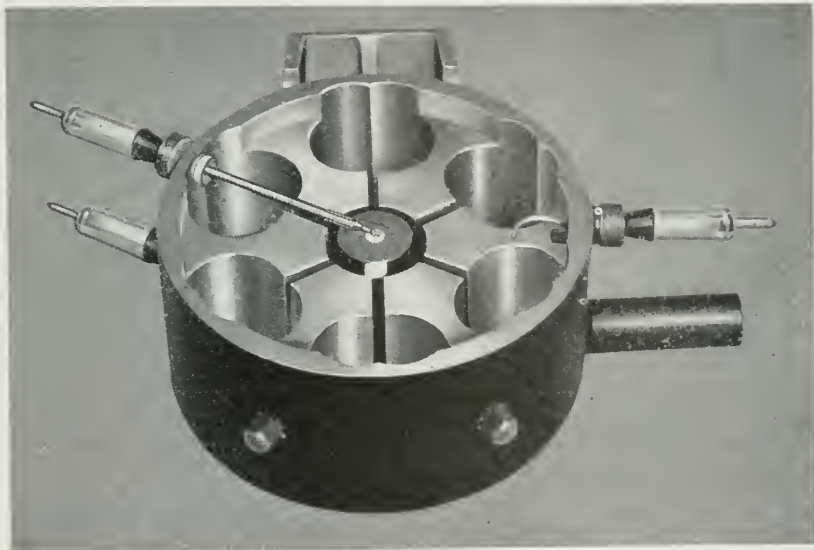
Fig. 29—W. E. 700-type magnetron—one side removed

and 1940. The first significant improvement came in England where work with multicavity magnetrons showed that this device was probably the answer to radar's need for a highly intermittent duty oscillator suitable for high power in the microwave region. A sample of this device was brought to this country by the Government and was tested in Bell Telephone Laboratories in October 1940. It produced pulses of several kilowatts at a frequency in the neighborhood of 3000 mc. A tremendous development of this device got under way immediately[5] and the multicavity magnetron

[5] "The Magnetron as a Generator of Centimeter Waves," J. B. Fisk, H. G. Hagstrum, and P. L. Hartman, *B. S. T. J.*, January, 1946.

became the key piece in the enormous development of radar equipment for still higher frequencies during the war. However, at the beginning of 1941 there were still many unsolved problems in 3000 mc radar other than that of the transmitter tube. On the other hand, the systems problems had been quite satisfactorily solved in the 700-mc region. The decision was therefore immediately made to extrapolate the British design down to
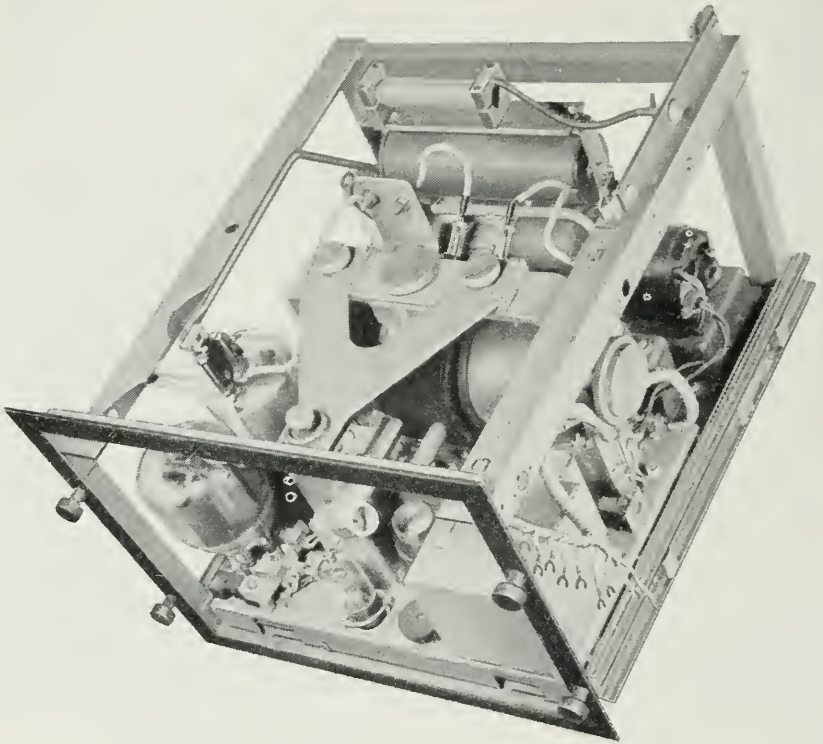
Fig. 30—Mark 3 & 4—transmitter

700 mc in order to obtain a higher powered and more satisfactory oscillator for the existing systems. This was done at top speed and a picture of the resulting tube is shown in Fig. 29. This was the first type of multi-cavity magnetron to go into production in this country. Concurrent with the design of the new magnetron the vacuum tube department of the Laboratories developed an improved tetrode modulator tube which was many times as efficient for radar pulse service as the triodes formerly used. This tube was designated W.E. 701-A.

A new transmitter using the magnetron and two of the new modulator

tubes was rushed through development and produced in time to go with the first accurate fire control radars. This transmitter provided a peak power output of about 40 kw with a pulse duration of 2 microseconds. It resulted in a material increase in reliable range, with satisfactory tube life. The new transmitter, shown in Fig. 30, was made mechanically interchangeable with the old and was applied retroactively also to the Mark 1 Radars.

Duplexing

The use of the high-power transmitter required additional protection for the receiver during the transmitted pulse in order to prevent damage
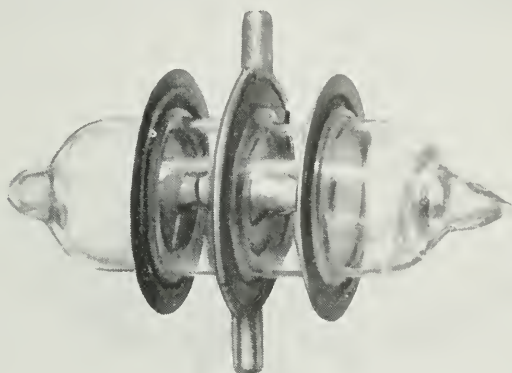
Fig. 31—W. E. 702—TR tube

and to permit the receiver to recover rapidly for reception of nearby echoes. The duplexing equipment was therefore modified to include a gas switching tube in the receiving transmission line. This was a refinement of the method used earlier by the Naval Research Laboratory.

The switching tube (W.E. 702A) was developed specifically for this purpose and is shown in Fig. 31. It was the first of the "TR" tubes of this general form and consists of a hydrogen-water vapor filled glass chamber with three copper electrodes.[6] This tube was mounted in the center of a half-wavelength coaxial line short circuited at each end, the outer conductor being connected to the outer electrodes and the center conductor

[6] "The Gas Discharge Transmit-Receive Switch," A. L. Samuel, J. W. Clark, and W. W. Mumford, this issue of *B. S. T. J.*

to the middle electrode. Input and output connections were tapped on this half-wave line near the short circuited ends. During reception this assembly introduces negligible loss in the receiving line. However, during the transmitted pulse a small amount of the transmitted power ionizes the gas in the switching tube and effectively short-circuits the receiver line. This device, which in later forms came to be called a "T-R Box", is located near the receiver input and the length of line between it and the junction with the transmitter line can be adjusted to an odd multiple of quarter wavelengths to present the desired high impedance at that point during transmission.

Receiver

The receiver delivered with early Mark 3 equipments was identical to that used in Radar Mark 1. It was of the superheterodyne type employing one stage of RF amplification (doorknob tube), 316A oscillator tube, and doorknob first detector. The intermediate frequency amplifier had a bandwidth of about 1 megacycle at a midband frequency of about 30 megacycles. The second detector and video stages were located in the indicating equipment. A photograph of this receiver is shown in Fig. 11.

Since in microwave work the controlling noise is that produced in the receiver, it is desirable to reduce this noise to the theoretical limit of thermal agitation in the input circuit. However, in 1939 tube limitations and circuit design techniques at these frequencies resulted in performance far short of this goal. The amount by which the receiver noise exceeds the theoretical minimum has been termed the receiver "noise figure" and in this early receiver the noise figure was about 24 db. It was recognized that considerable improvement in maximum range could be obtained by reducing this receiver noise.

Shortly after first deliveries of Radar Mark 3 a new tube (GL-446 or "lighthouse" tube) was made available by the General Electric Company which showed promise of providing a substantial improvement in the receiver noise figure. An amplifier using this tube was accordingly designed by Bell Laboratories in which coaxial cavities were used for tuning elements. Two stages of amplification were used to replace the single "doorknob" tube stage previously employed. The new amplifier resulted in a reduction of the receiver noise figure to about 9 db and provided a marked improvement in maximum range capability of the radar. These amplifiers were manufactured and shipped to the Fleet for field installations on early equipments and were included in productions on equipments shipped subsequently to availability of the amplifiers. A photograph of the receiver with the two amplifiers installed is shown in Fig. 32.

Another field modification provided automatic gain control of the signal
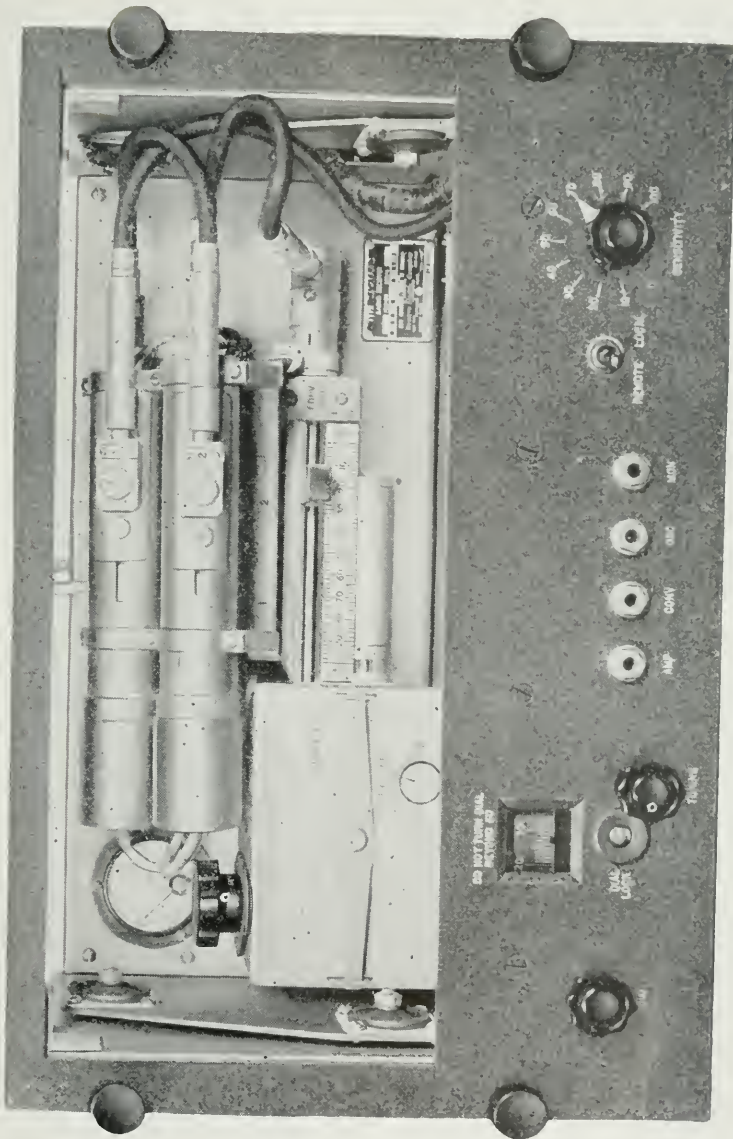
Fig. 32—Receiver—showing improved R. F. amplifier

selected by the range operator. This was supplied in the form of an external unit which controlled the gain of the receiver IF amplifier to reduce signal fluctuations produced by fading.

The first production Mark 3 Radars were delivered to the Navy in October 1941, and the first two installations were completed on the main battery directors of the U.S.S. Philadelphia at the Brooklyn Navy Yard that month.

RADAR MARK IV

During the development work on Radar Mark 3 the Navy pointed out the need for a fire control radar for use with the 5-inch Naval guns against enemy aircraft. The Bell Telephone Laboratories was therefore requested to further modify the radar design to meet this need. The anti-aircraft equipment was first designated FD, later becoming known as Radar Mark 4.

For antiaircraft fire control a new coordinate had to be added to the target-locating system; namely, elevation angle. Again it was desired that the additional information be obtained from the single antenna with a precision equal to that already obtained in azimuth. This problem was approached in a manner similar to that used for the Mark 3 antenna and is described below.

Two Plane Lobe Switching

In considering two plane lobe switching methods it appeared that the desired result could be obtained by mounting two 3 ft. x 6 ft. parabolic arrays one above the other. This arrangement was tried and resulted in the array shown in Fig. 33. It provided two plane lobe switching with an antenna only slightly larger than the 6 ft. x 6 ft. antenna used before and had comparable gain and beam width (see Table II).

A schematic diagram of the array is shown in Fig. 34. Here it will be seen that there are two horizontal dipole arrays, each mounted along the focal line of a cylindrical parabola. The dipoles are in four groups and the interconnecting harness is criss-crossed and joined to the feed line at the center. Symmetrically placed around the feed point are four stub lines connected to the lobe switcher stators. Here again a semi-circular rotor is used for the lobing shifting capacitor. It will be observed that during each quarter turn of the rotor two stator plates are engaged, and the sequence is such that the beam shifts left, up, right, and down during one rotation. A separate Indicator was provided for the Pointer (elevation operator). To avoid signal confusion on the two Indicators it is necessary to show only left-right signals on the Trainer's oscilloscope and up-down signals on the Pointer's oscilloscope. This is accomplished by means of cam operated contacts in the lobe switcher which blank the indicators during the required

intervals. Other contacts on this assembly provide left-right and up-down image spacing.
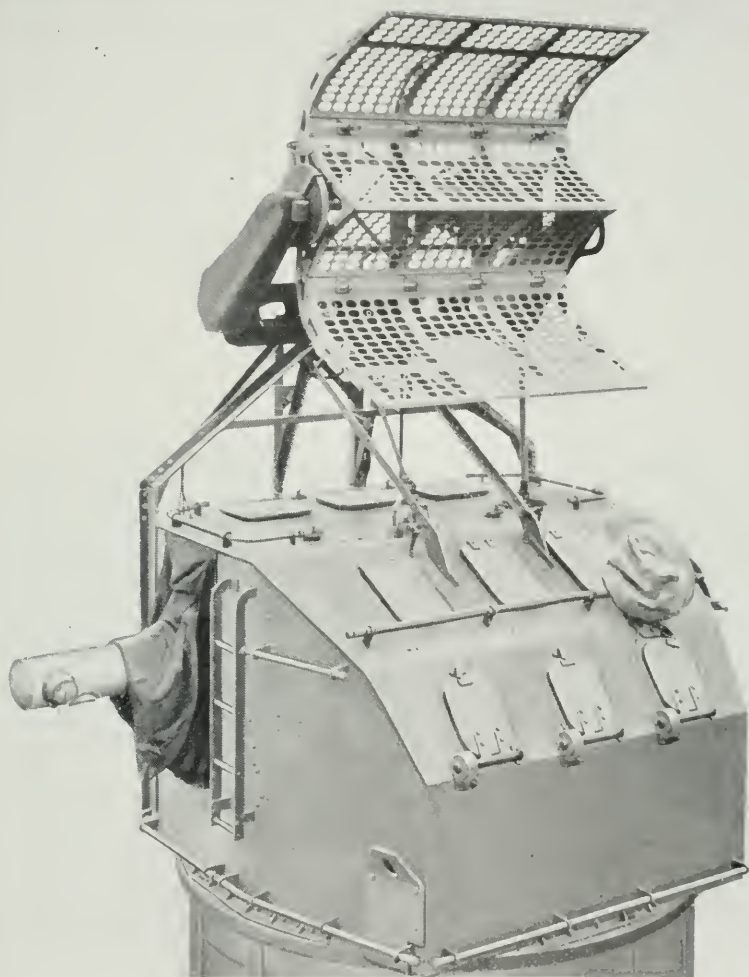
Fig. 33—Mark 4—antenna on gun director

Except for the new antenna and the additional Train or Elevation Indicator for the Pointer (elevation operator), this radar was identical to Radar Mark 3. The first demonstration of a development model of Radar Mark 4 was made at Atlantic Highlands, New Jersey, in September 1941 and this model was installed aboard the destroyer U.S.S. Roe the latter
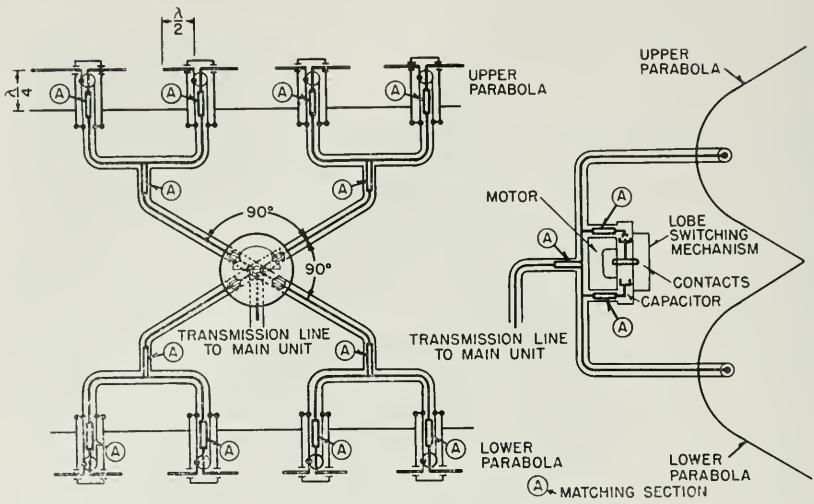
Fig. 34—Mark 4—Antenna schematic
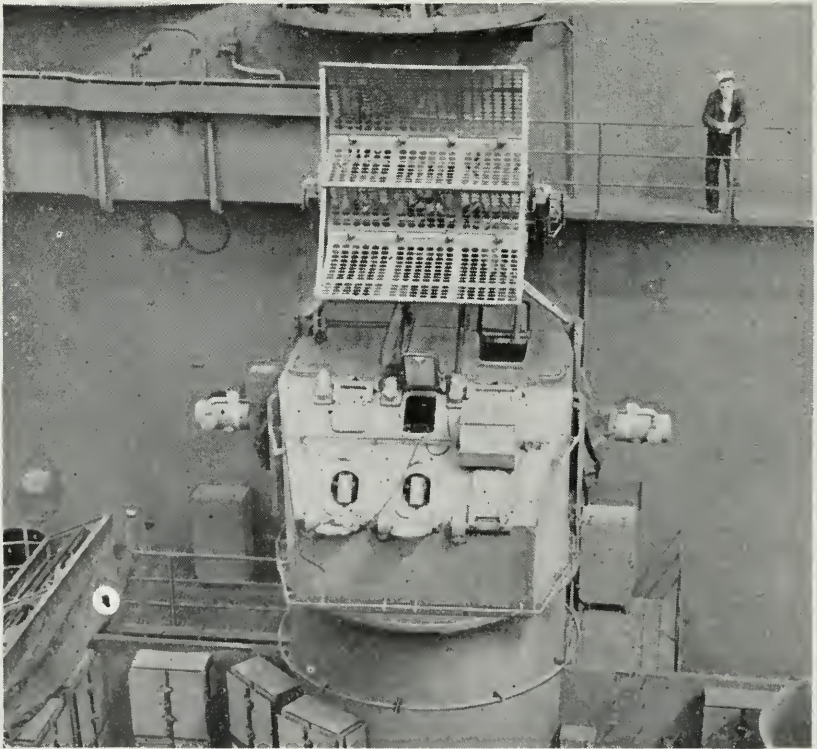
Fig. 35—Mark 4 antenna on Battleship Tennessee (Navy Photo 1908-43)
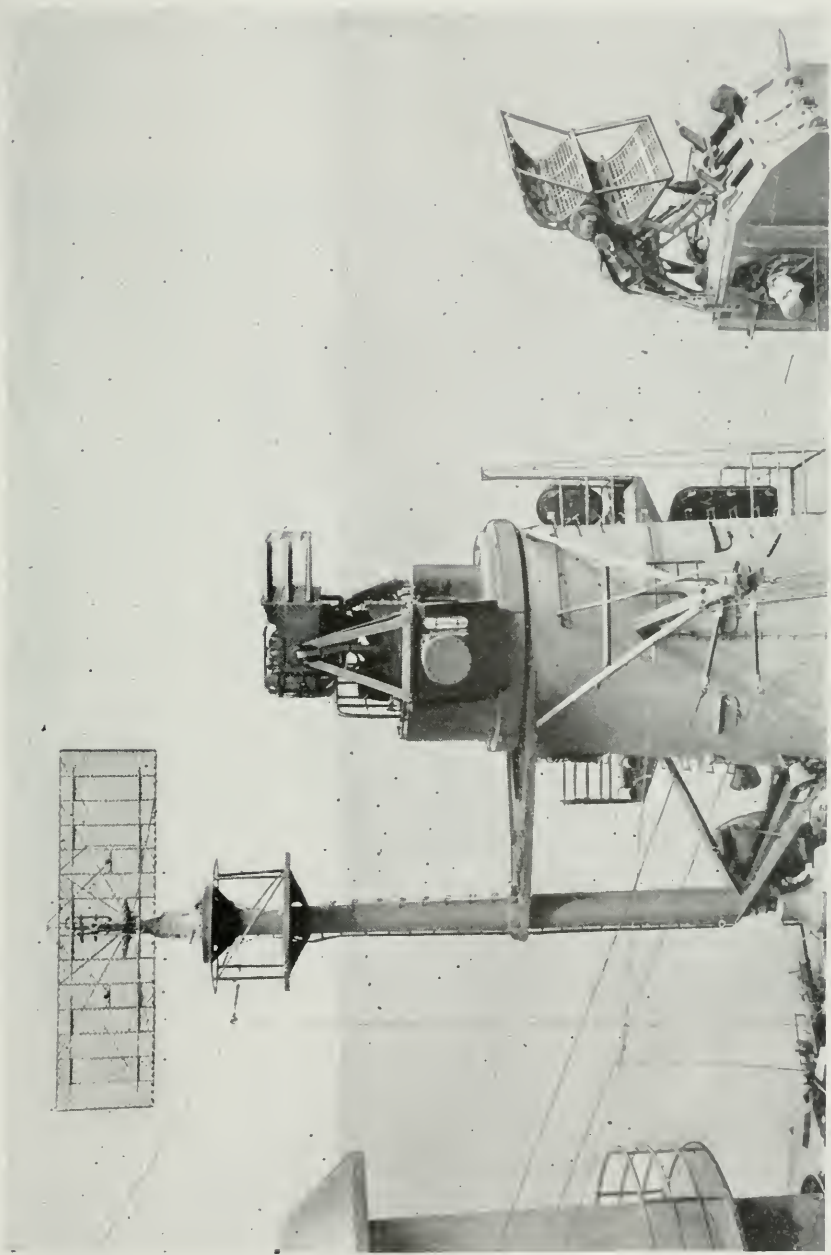
Fig. 36- Radar antennas on Battleship Tennessee (Navy Photo 1905-43)

part of that month.   Initial production deliveries of these radars were made
in December 1941.

Typical installations of the Mark 4 Antenna on the secondary battery
directors of a battleship are shown in Figs. 35 and 36.   The main frame
installation for Mark 3 and Mark 4 on a battleship is shown in Fig. 37

Fig. 37—Radars Mark 3 & 4—main units on Battleship New Jersey (Navy Photo 181809)

while typical installations of the train and elevation operator's units in the
director are shown in Figs. 38 and 39.

Application and Use of Mark 3 and 4 Radars

The Mark 3 radars, designed for use against surface targets only, were
generally installed on the main battery directors of battleships and cruisers.
The Mark 4 radars for use against either surface targets or aircraft were
generally installed on the secondary battery directors of battleships and
cruisers, and on the one and only dual purpose director on destroyers.
Thus a battleship usually had two Mark 3 and four Mark 4 equipments and
a destroyer one Mark 4.   Practically every ship in the fleet, of destroyer

size or larger, was equipped with one or more of these equipments early in
the war. A total of 139 Mark 3, and 670 Mark 4 radars were built, in-
cluding those used ashore at schools. Although some of these equipments
were replaced by more modern designs before the end of the war and some
were lost in battle, there were still approximately 85 Mark 3 and 300 Mark 4
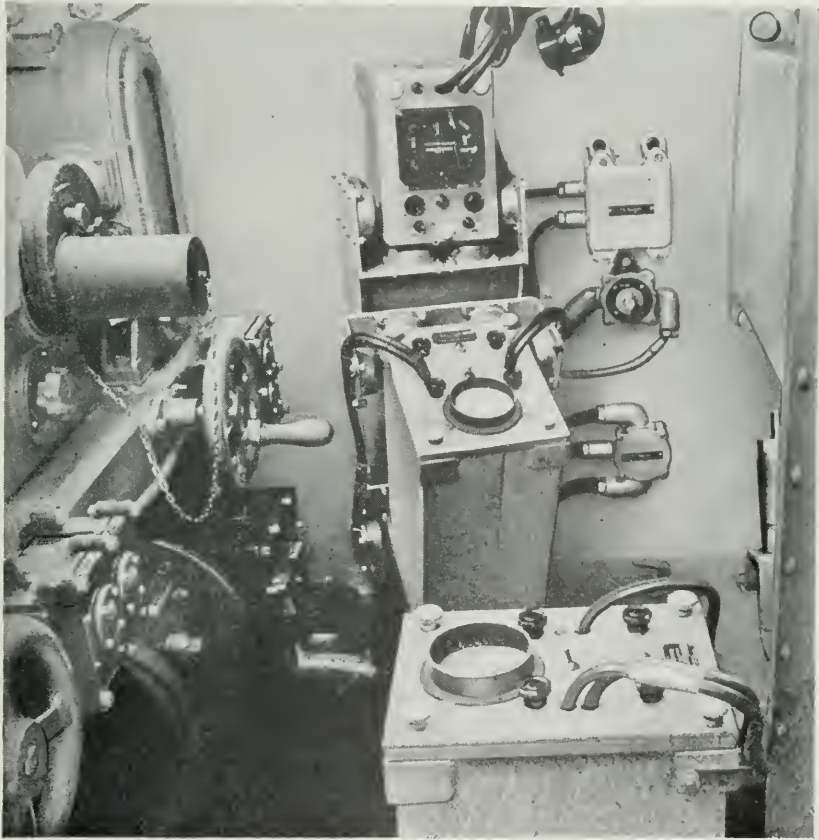
Fig. 38—Mark 4 Radar—trainer & pointer operators' positions on Aircraft Carrier Saratoga
(Navy Photo 177347)

radars in service in the fleet on V-J day. The first four Mark 4's, Serial Nos.
1, 2, 3 and 5 installed on the battleship Washington were used until the
middle of 1945, although newer designs had been going on all new vessels
for more than a year.

These early equipments were the "guinea pigs" of fire control radar.
They were the instruments with which our fleet learned to fight effectively

at night and thereby gain a large advantage over the enemy whose radar was feeble and inaccurate. They played a part in every one of the early battles and most of the later ones in the Pacific. They controlled the cruiser Boise's guns in October 1942, when she blazed away at night at a vastly superior fleet in the Solomons and made the enemy pay 10 to 1 for the
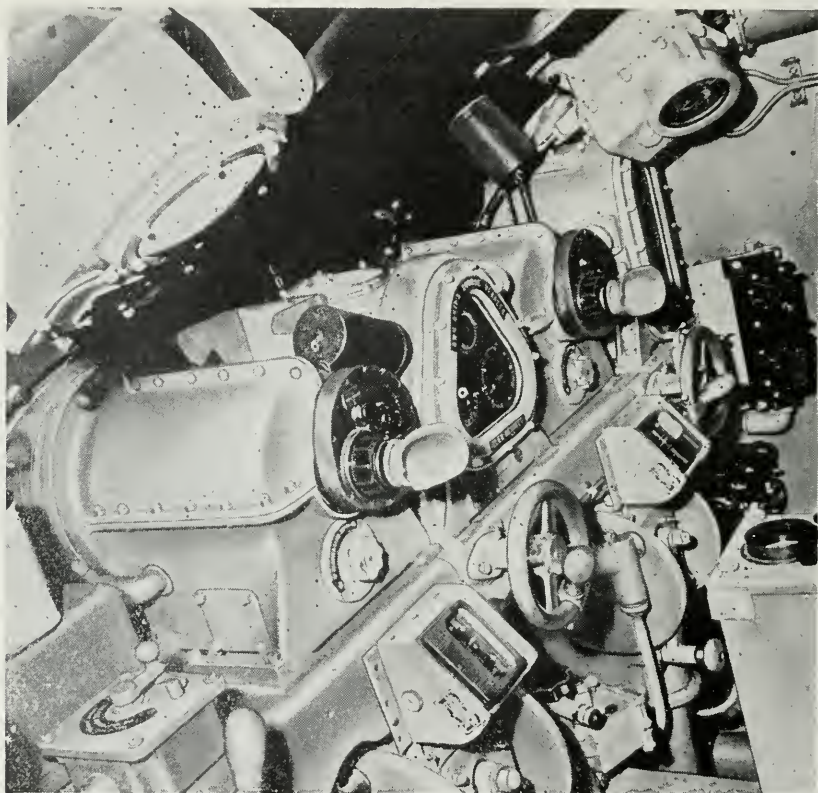
Fig. 39—Radar Mark 4—trainer & pointer operators' positions on Destroyer Barton
(Navy Photo 181775)

damage they succeeded in doing. They were with the cruiser San Francisco on a night in November 1942, when a small U. S. force sank 27 enemy ships, almost completely destroying a large Japanese convoy bound for Guadalcanal when our hold there was at best precarious. The Mark 3 steered the big guns of the battleship South Dakota in the Solomons on the dark night of November 4, 1942, when she sank a major Japanese war vessel eight miles away with two salvos. Even in engagements in broad daylight when optics could be used for target angles these radars still played a vital

part in furnishing accurate range which made 5″ gunfire against aircraft, for example, deadly at long range. Thus on October 16, 1942, when the South Dakota was attacked by planes she shot down an even 38 out of 38 attacking.

The rapid and widespread application of this rather complex electronic equipment was not accomplished without pain and confusion. It is beyond the scope of this paper to discuss the enormous problem of training in operation and maintenance that had to be solved, or of the tactical revolution in Naval warfare that fire-control radar produced. It is sufficient here to say that these and other problems were solved by heroic efforts of hundreds of officers and civilians in the Navy Department ashore and the thousands of officers and men of the fleet. Their problems were made more difficult by weaknesses in the equipment which were revealed by battle experience as the new science of radar got its baptism of fire. In every possible case the Laboratories attempted to remove the causes of recurring troubles by redesign and the furnishing of improvement kits of parts for installation in the fleet. The many lessons of experience learned from the Mark 3's and 4's were immediately applied in the design of the many more modern radars for the same and other types of service.

The authors of this paper wish to express their gratitude to the many Navy men with whom they have worked in connection with these equipments, and whose whole-hearted cooperation during difficult times made possible the successful development of these fire-control radars. They also wish to thank their colleagues in Bell Telephone Laboratories who worked as a team to make this important equipment possible, and the men of the Western Electric Company for their help on the many engineering problems which arose during production and use in the field. It is the hope of all who were concerned with this development that accurate radars, like other radars, will find peaceful use in a peaceful world, but it is also the determination of these engineers that as long as we need a Navy, we will try to provide it with radars as much superior to those of any possible enemy as they were in the recent war.

The Gas-Discharge Transmit-Receive Switch

By A. L. SAMUEL, J. W. CLARK and W. W. MUMFORD

THE gas-discharge transmit-receive switch has become an accepted part of every modern radar set. Indeed, without such a device, an efficient single-antenna micro-wave radar would be nearly impossible. Many of the early radar sets made in this country employed separate antennae for the transmitter and receiver. The advantages of single antenna operation are so apparent as hardly to require discussion. The saving in space or, if the same space is to be occupied, the increase in gain and directivity of a large single antenna is, of course, apparent. But even more important, perhaps, is the tremendous simplification in tracking offered by a single antenna, particularly where a very rapid complex scanning motion is desired.

The fact that the receiver needs to be operative only during periods between the transmitting pulses makes single antenna operation possible if four conditions are satisfied. These are: (1) the receiver must not absorb too large a fraction of the transmitter power during the transmitting period, (2) the receiver must not be permanently damaged by that portion of the transmitter power which it does absorb, (3) the receiver must recover its sensitivity after any possible overload during the transmitting pulse in a time interval shorter than the interval required by the reflected pulse to arrive back to the receiver from the nearest target, and (4) the transmitter must not absorb too large a fraction of the received power. At frequencies of the order of 700 megacycles and at low power levels these conditions are not impossible of attainment without recourse to any special switching mechanism other than that provided automatically by the usual circuit components. Conditions (1) and (2) can be met by designing the receiver in such a way that the change in input impedance as a result of overload will cause most of the available input power to the receiver to be reflected. Condition (3) requires careful attention to the time constants of all those receiver circuits which are subject to overload. Condition (4) fortunately is automatically satisfied by most transmitters, again as a result of the large mismatch reflections which occur at the connections to the transmitter's "tank" circuit when the transmitter is not operating. The United States Navy Mark 1 radar was operated on this basis.

The speed with which the transmit-receive switch must operate rules out all consideration of mechanical devices, at least for all but the longest range

"early warning" equipment. For example, the go and return time to a target at 500 feet distance requires approximately one microsecond. Switching times must, therefore, be measured in microseconds. Since these short time intervals would at first sight seem to be too small to permit the use of gaseous discharge devices, some work was done on the use of specially designed vacuum diodes. It is possible to employ balancing circuits (sometimes called hybrid circuits) to achieve single antenna operation, but such circuits require critical balancing adjustments and they dissipate a large part of the available power in non-useful balancing loads. The need for a still different approach to the duplexing problem was clearly indicated.

Spark discharges either in air or in enclosed gaps bridged across parallel wire transmission lines were used in some of the early experimental long-wave radar sets. Dr. Robert M. Page of the Naval Research Laboratory was one of the pioneers in this work. These devices were only moderately satisfactory because of their erratic behavior and because of electrode wear. However, it was observed that the recovery time of such discharges was not as long as might be expected on the basis of a simple ionization and deionization explanation of their operation. This led to the investigation of the use of low-pressure gas discharges. These very early gas-discharge "switches" were actually much more in the nature of "lightning protectors", their principal function being to limit the power delivered to the receiver during the transmitting pulse in a gross sort of way, with considerable reliance on impedance changes at the receiver and on the rugged overload capabilities of the first tube in the receiver.

The trend toward shorter wavelengths and the desire for better protection led to the development of a partially evacuated gas-discharge tube located in a relatively high Q resonant cavity. In England, cavity type duplex tubes were made by inserting gas in a then current type (Sutton Tube) of local oscillator tube. These devices were called TR boxes (abbreviation for transmit-receive) by the English, a designation which has continued. It is a curious coincidence that some of the earliest cavity type duplex tubes made in this country at the Bell Telephone Laboratories were also constructed by inserting gas in an American type local oscillator tube (the 712A vacuum tube). This tube (later coded the 709A vacuum tube) was tested in an operative system which was subsequently demonstrated to the Army with such satisfactory results that the tube was adopted without change for several radar systems. The 709A vacuum tube and its associated cavity are shown in Fig. 1.

A similar structure, known as the 702A and shown in Fig. 2 (together with the 709A tube) was used for longer wavelengths. The need for these tubes was so very great that no time was allowed for their improvement before production was undertaken by the Western Electric Company.

709A VACUUM TUBE

INPUT COUPLING LOOP

CAVITY

OUTPUT COUPLING LOOP

TUNING PLUG
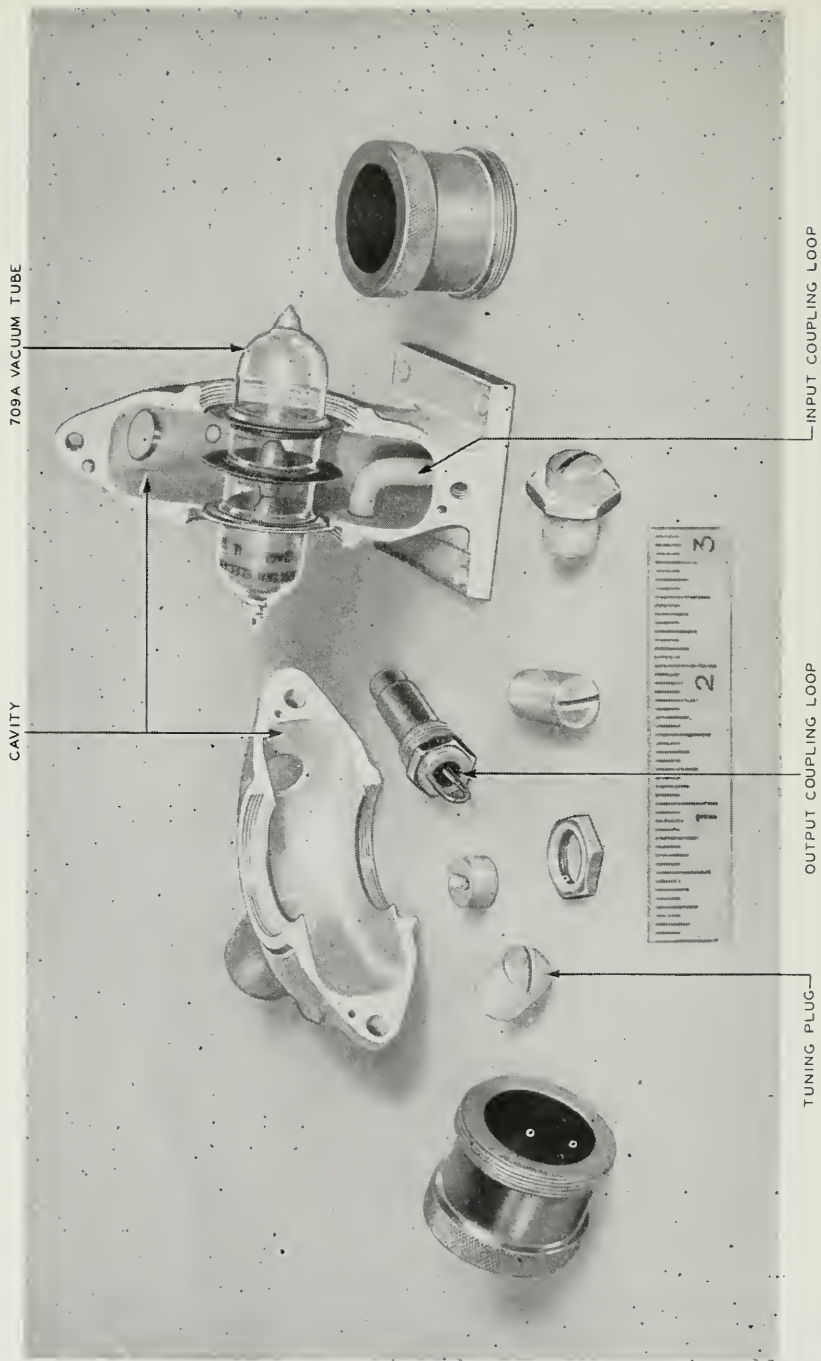
Fig. 1—The 709A vacuum tube and its associated cavity

The radar on the U.S.S. Boise in the battle off Savo Island on October 11–12, 1942 employed a 702A vacuum tube.

Three developments soon led to the need for much improved TR boxes. One of these was the rapid progress which was being made in increasing the peak output power from the magnetron. The second was improvements in the silicon point contact rectifier, the so-called crystal detector, which increased its reliability and convenience and at the same time reduced its conversion loss and noise figure as compared with the vacuum tube converter. The third was the development of still higher frequency systems to
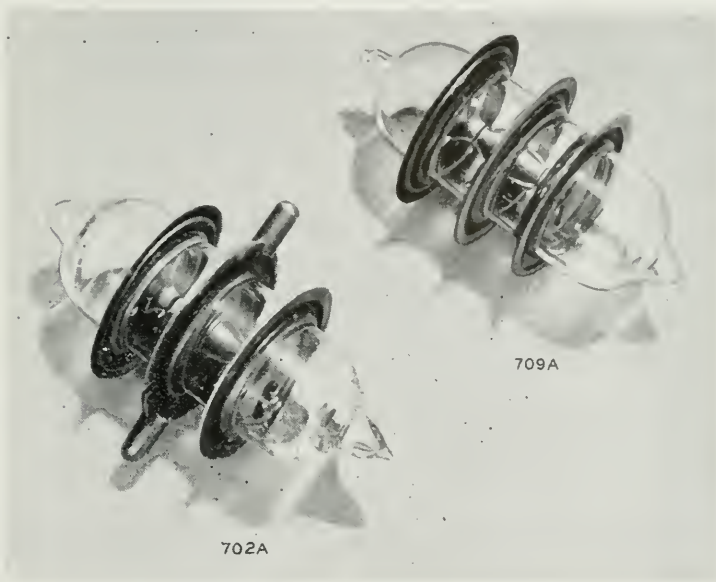
Fig. 2—The 702A and 709A vacuum tubes

achieve either greater antenna directivity or smaller size. Since satisfactory vacuum tube converters were not available for these frequencies, the silicon rectifier had to be used. Unfortunately the silicon rectifier, as then available, was subject to permanent damage if subjected to but very small amounts of power as compared with the magnetron power levels.

An active program of work was initiated at the Bell Telephone Laboratories to obtain designs of TR boxes offering adequate protection for contact rectifiers at any power levels then available or contemplated. Three tubes were developed, the 721A, 724B and 1B23 vacuum tubes shown in Fig. 3. These tubes are used at frequencies in the vicinity of 3000 megacycles, 10,000 megacycles, and 1000 megacycles respectively. They are all of the
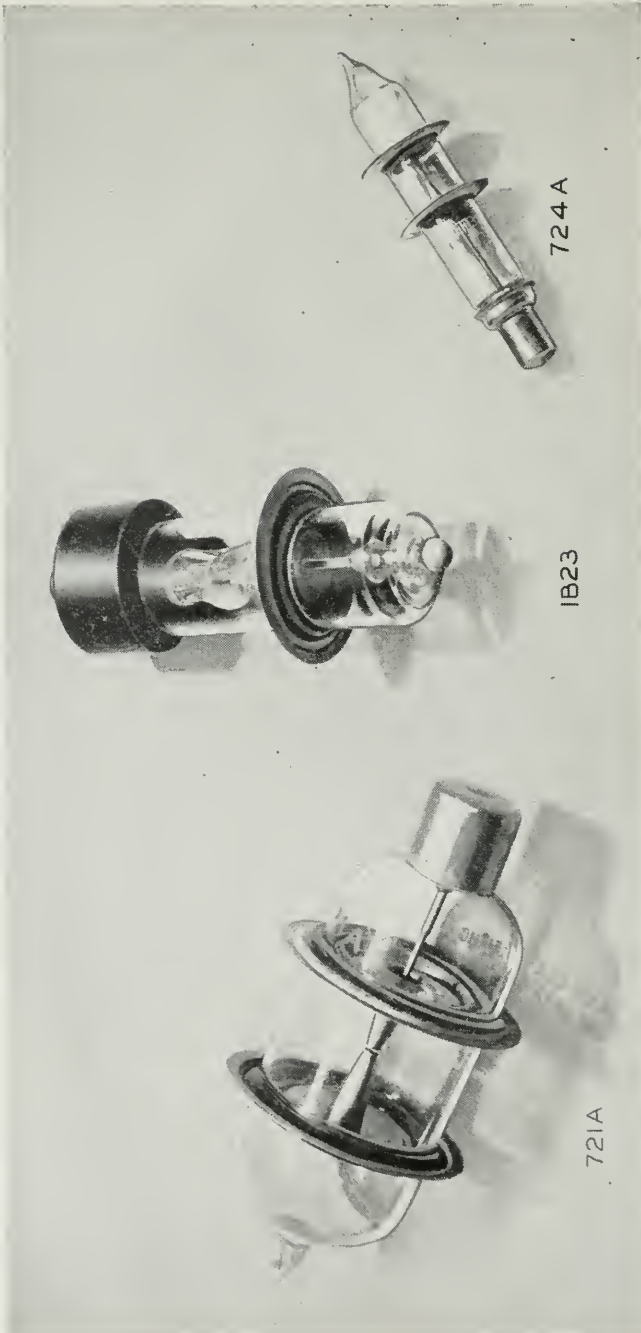
Fig. 3—The 721A, 724B and 1B23 vacuum tube

separate cavity type in which contact through the vacuum envelope is made by means of thin copper discs. More recently other designs of tubes have appeared in which the entire cavity is evacuated. Over 400,000 tubes of the three types discussed in this paper were manufactured in 1944 alone and substantially all of the American-made radars which saw active service employed one or more of these tubes.

The TR tubes used in American radars are, of course, no more essential than are the magnetrons, the beating oscillators, and the many other special parts which go to make up the modern radar. Nevertheless it is interesting to note that the 721A tube was an essential part of the radar equipment on almost every major ship in the United States Fleet, that the 724B tube was an essential part of the bombing equipment on nearly every bomber used against Japan, including the planes which carried the atomic bombs, and that the capture of Okinawa, to name a single case, would have been much more expensive in men's lives without equipment depending upon the 1B23 tube.

Method of Operation

The 709A tube as shown in Fig. 1 was operated in what has come to be known as a shunt branching circuit. Its operation can be explained in terms of Fig. 4. During transmission, energy flows from the transmitter along the coaxial line toward the antenna. Some of this energy enters the branch leading to the receiver where it encounters the TR box. This consists of a resonant cavity with a pair of spark gap electrodes arranged so that the maximum resonant voltage is built up across the gap. Since the voltage across the gap is then limited by the discharge voltage and the voltage applied to the receiver is still further reduced by an equivalent step-down ratio of the output coupling in the resonant cavity, the receiver input power is held to a small value. The power dissipated in the gas discharge, and therefore abstracted from the transmitted signal is kept small by the impedance mismatch. The discharge itself takes the form of a small pale blue glow between the electrodes. The effect of the discharge is to place a low impedance (predominantly resistive) across the maximum impedance point of the cavity. This results in the appearance of a still lower apparent impedance across the input to the cavity. If the length L_1 is an odd number of quarter wavelengths, the apparent impedance of the receiver branch at the branching point becomes very high in comparison with the impedance of the antenna and is therefore unable to abstract much power from the line.

At the end of the transmitting period, the conductance of the gas discharge falls rapidly to a very low value since the small received voltages will be insufficient to maintain the discharge. Signals arriving at the antenna can then be transmitted through the TR box to the receiver. However,

in the circuit shown in Fig. 4, the receiver is still bridged by the transmitter. It is a fortunate fact that the internal impedance of many magnetrons (the most common type of transmitting tube) becomes very low when they are in the inoperative condition so that the tube is nearly the equivalent of a short circuit. By adjusting the length L_2, until this equivalent short circuit position is an odd number of quarter wavelengths from the junction point "B", the shunting impedance at "B" can be made very high so that only a small part of the received energy is lost. In the event that this change of impedance of the transmitter is not sufficient, a second TR switch commonly known as an ATR, may be introduced to perform this function as will be described later. During the receiving period, some loss will occur in the TR box resonant cavity as a result of the inherent resistive and dielectric losses. An additional loss will occur immediately after the
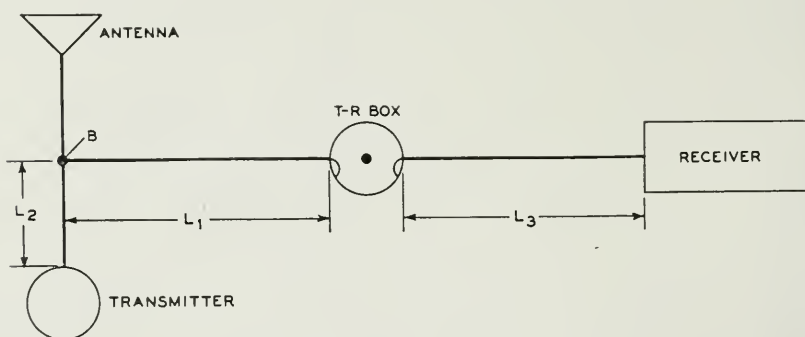
Fig. 4—The elements of a shunt branching circuit

transmitting period because of the loss producing particles (free electrons) which remain for a time in the discharge gap. The combined losses must be kept small so as not to impair the performance of the system.

Most modern radars employ series branching circuits instead of the shunt branching circuit just described. A coaxial line example of such a system (from the SCR-545) employing the 721A tube is shown in Fig. 5. As shown in Fig. 6 the cavity is coupled to the coaxial line by means of a window which can be slid along on a slot in the outer conductor of the coaxial line leading from the transmitter to the antenna. Fig. 7 is an exploded view of the cavity. Such a cavity is in effect in series with the line as the currents existing in the outer conductor of the coaxial line are interrupted by the window. During the transmitting period the low impedance at this window limits the voltage across it to a small value and prevents serious loss of transmitter power.

Reception in the series branching circuit of Fig. 5 is achieved by adjusting
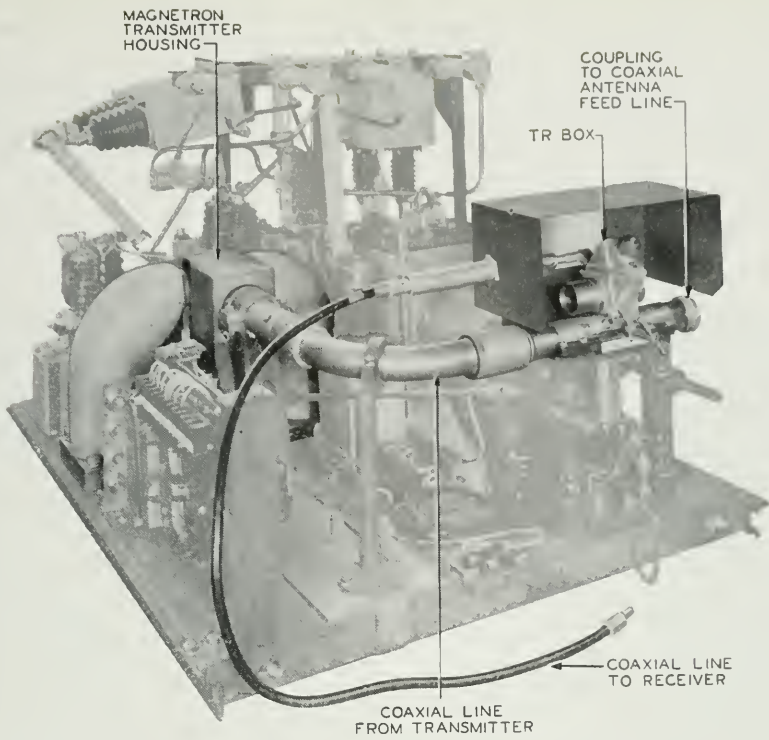
Fig. 5—The series branching circuit employing the 721A tube used in the SCR545
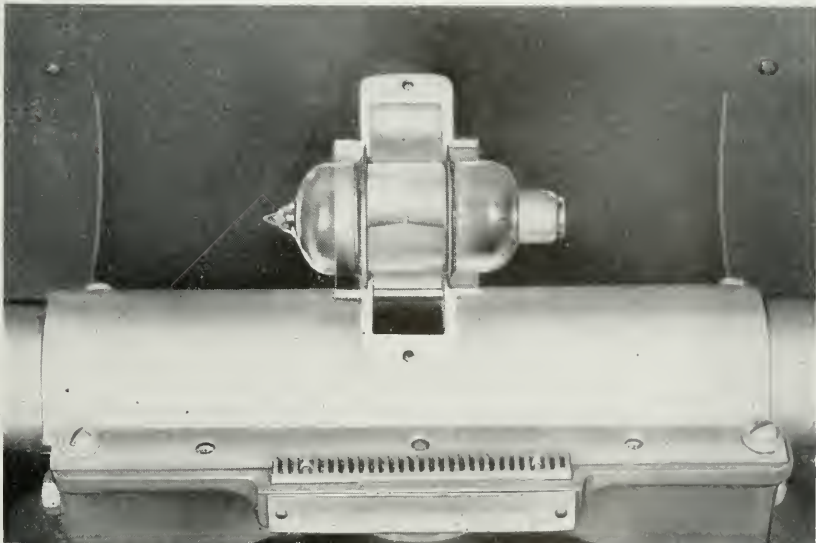
Fig. 6—Closeup view of the cavity shown in Fig. 5, partly disassembled
to show coupling window

the position of the TR cavity along the slotted section of the line until the window is located the correct distance from the transmitting tube. This now corresponds to an even number of quarter wavelengths between the equivalent short-circuit plane and the junction, so that the maximum current is caused to enter the cavity. The output to the receiver is obtained from a small coupling loop in the TR cavity.
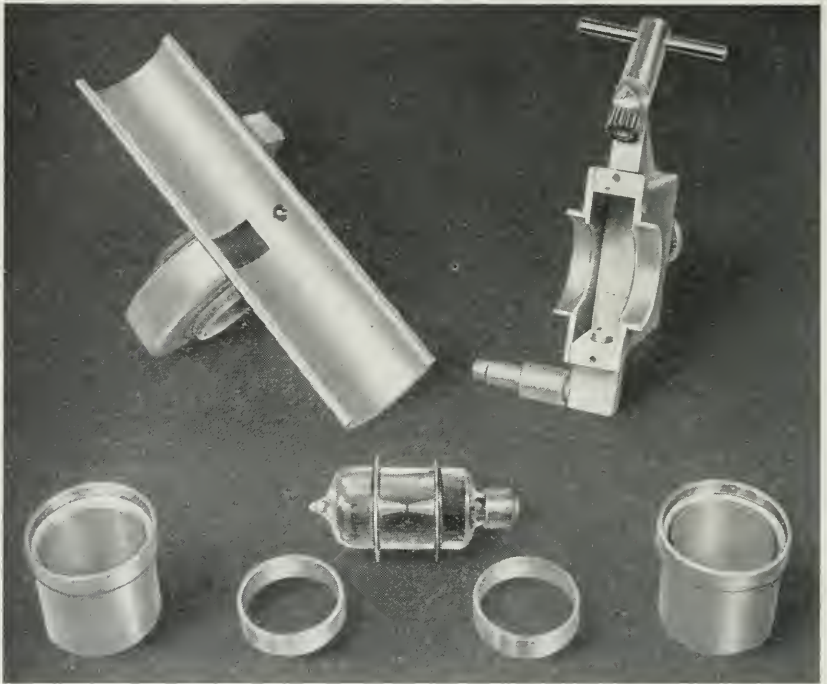
Fig. 7—Exploded view of the cavity of Fig. 6

A similar coaxial line series branching circuit using the 1B23 is shown in Fig. 8. The method of inserting the tube in the cavity is shown in Fig. 9 while Fig. 10 is an exploded view showing the details.

This method of coupling a resonant cavity to a transmission line by a window is not limited to the coaxial line case. A wave guide system is shown in Fig. 11. The distance between the TR cavity and the transmitting tube is again adjusted by sliding the cavity and its window along over a section of the rectangular wave guide containing a slot.

The ATR. In all of the systems so far described use is made of the impedance mismatch conditions at the magnetron or other transmitting tube terminals to prevent serious loss during the receiving period. If the

magnetron "cold" impedance does not differ greatly from the surge imped-
ance of the transmission line used, it may not be possible to avoid loss of
reflected signal into the transmitter line.   Also in some cases, an unreason-
able amount of adjustment must be provided in the position of the TR
cavity with respect to the transmitter to make up for large variations which
may be encountered in transmitter "cold" impedance.   Both difficulties
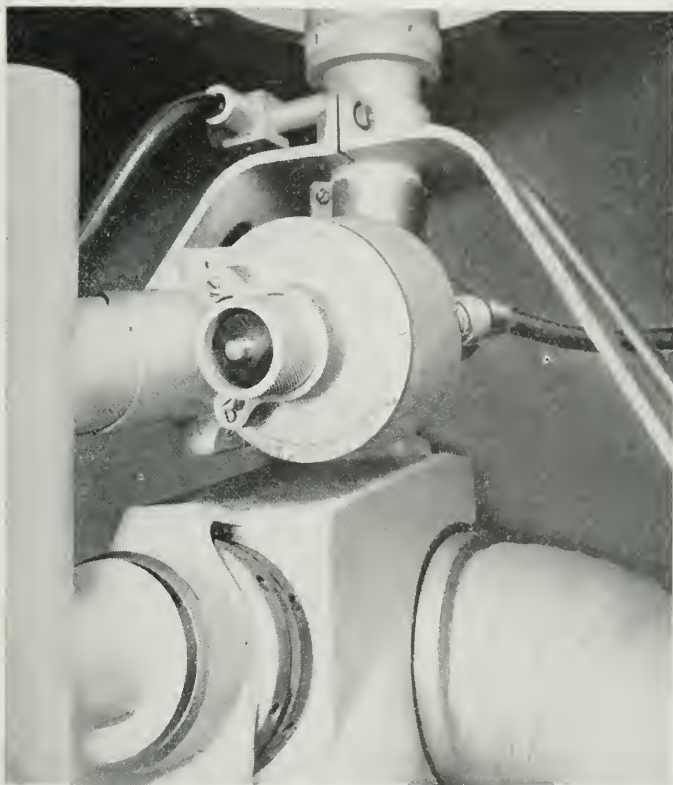
Fig. 8—Series branching circuit using the 1B23 vacuum tube

may be avoided by the use of a second gas discharge tube located between
the transmitter and the TR and at an odd number of quarter wavelengths
from the TR junction.   This second tube is referred to as the anti-T-R tube
(usually abbreviated to ATR), or sometimes as the RT tube.   The use of
an ATR tube was not found to be necessary in most of the systems which
employed the 721A tube.

  With the advent of still higher frequency systems, for which the 724B
tube was designed, the "cold" impedance difficulties just mentioned made

TR BOX
CLAMPING
SCREWS

TR BOX

VACUUM
TUBE
1B23

KNURLED
TUBE
CLAMPING
SLEEVE
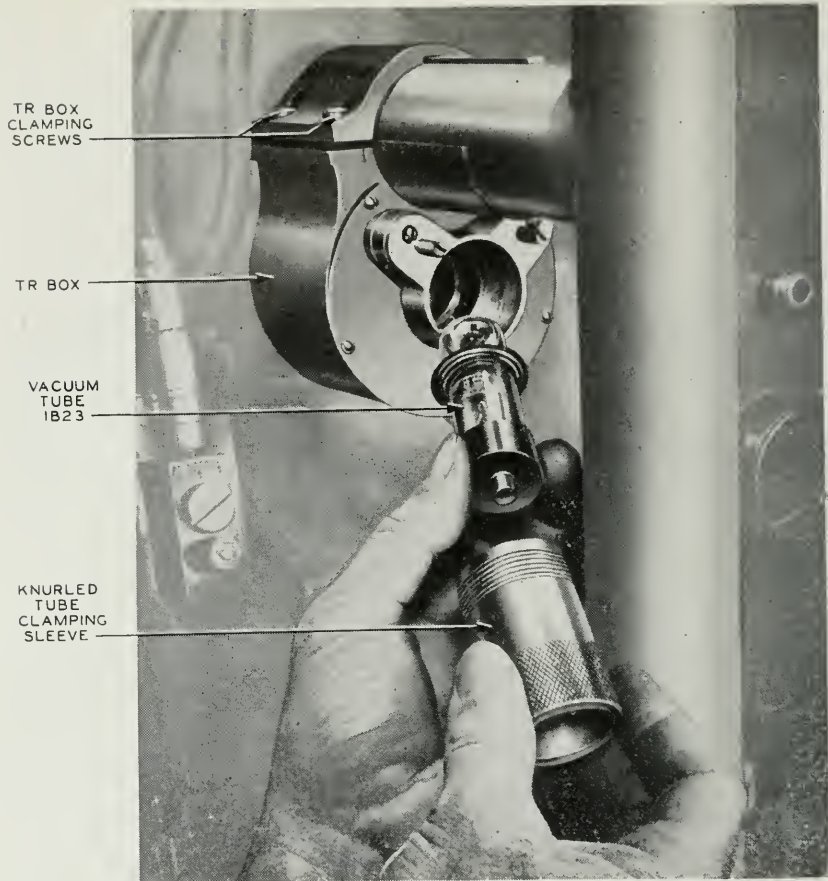
Fig. 9—Method of inserting the 1B23 into the circuit

TR BOX

OUTPUT
CONNECTION

LOCKING SCREW

TUNING SCREW

VACUUM TUBE

TUBE HOLDER

CLAMP

TUNING PLATE

CLAMP SCREWS
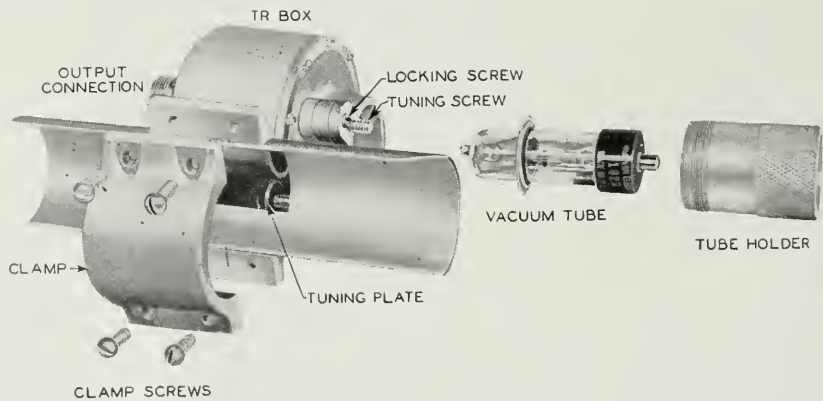
Fig. 10—Exploded view of the 1B23 cavity

it seem advisable to employ an ATR tube. The general arrangements of the circuit elements in one of these systems is shown in Fig. 12. The main wave guide section is shown removed from the rest of the equipment
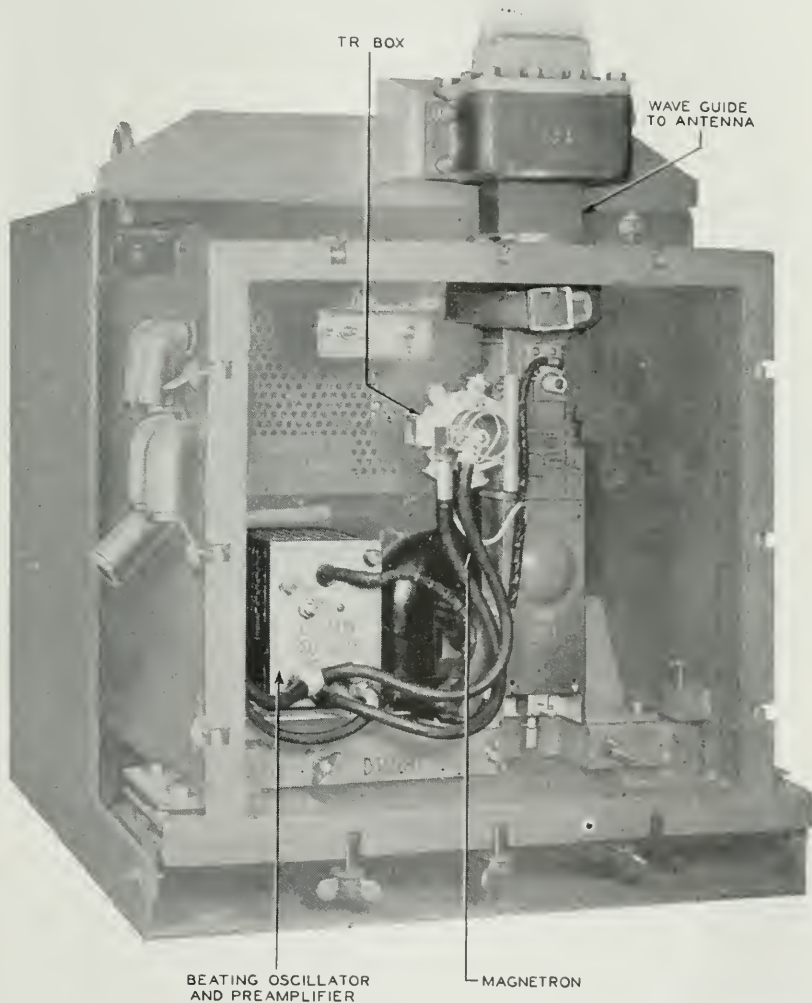
Fig. 11—The wave guide system of the SL radar employing the 721A tube

in Fig. 13 while Fig. 14 is an exploded view revealing the details of the cavity construction. The two cavities are, of course, coupled to the wave guide by means of windows. The wave guide branch leading to the receiver is in
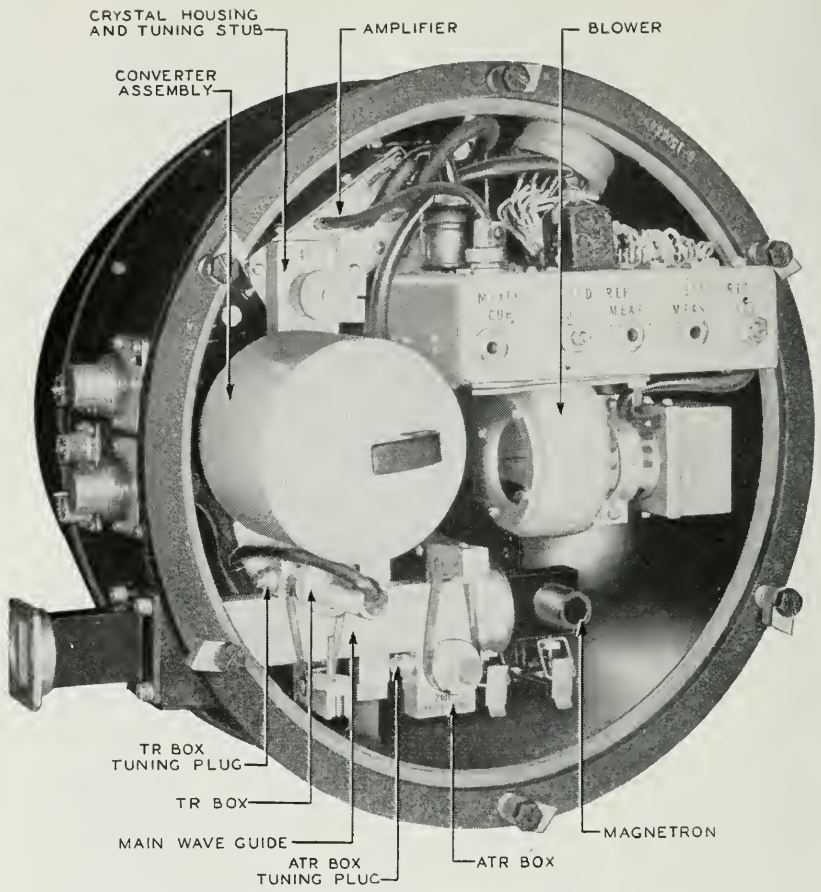
Fig. 12—A general view of a wave guide system employing a 724B tube in the TR box and a second 724B tube in the ATR
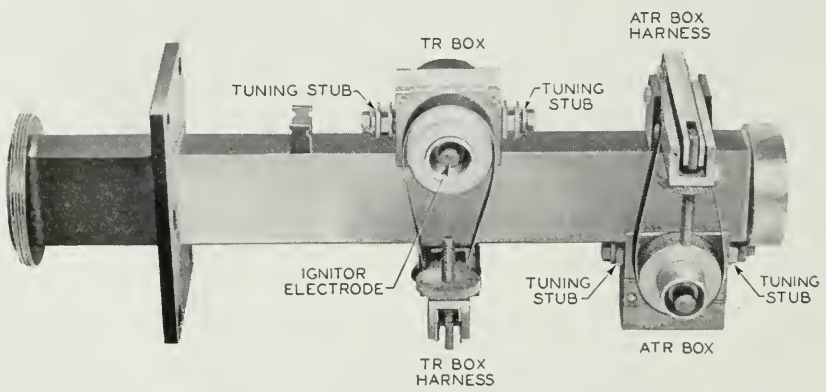
Fig. 13—The main wave guide of Fig. 12 removed from the rest of the equipment

turn coupled to the TR cavity by another window. The input and output windows of the TR cavity are adjusted in size to provide an impedance match to the line during the receiving period. The window to the ATR is, however, adjusted so that it presents a high impedance, that is much greater than the surge impedance, during the receiving period. This high impedance is effectively in series with the magnetron impedance. The resulting high impedance is located at an odd number of quarter wavelengths from the TR and so presents a very low impedance in series with the receiving branch.
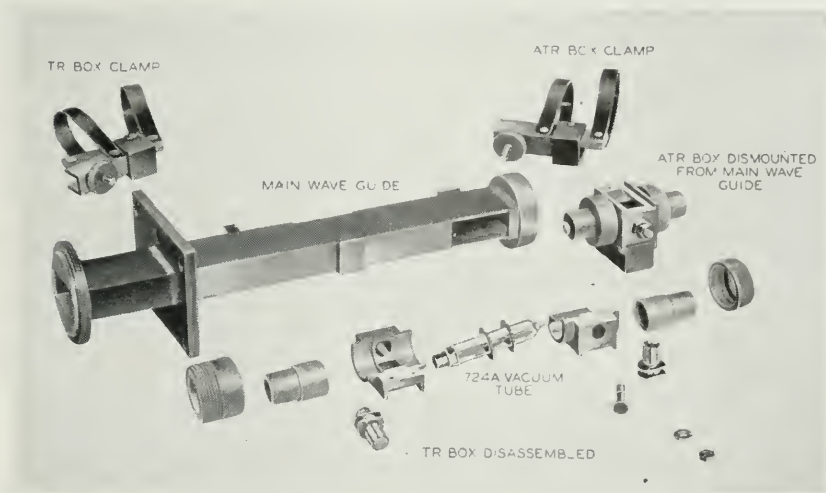
Fig. 14—An exploded view of Fig. 13

Both the TR box and the ATR box must be tuned to resonance at the magnetron frequency. Broad-band ATR boxes using very low Q circuits have been designed which require no adjustment over a 5% band. Such boxes, which obviously are very advantageous in tunable systems, do not use the copper-disc-seal tubes which form the principal subject matter of the present paper, and will not be discussed further.

TR Box Performance

The performance of a TR box can be described in terms of four parameters which are related to the four duplexing requirements mentioned earlier. These parameters are respectively: (1) the high level loss, which is the transmitting power loss in the TR tube; (2) the leakage power, which is the amount of transmitter power which reaches the receiver; (3) the recovery time, which measures the rate at which the TR box recovers its low level

behavior after the termination of the transmitting period; and (4) the low-level loss, which describes the loss of the received signal including (a) the loss in the TR box itself and (b) the loss in the transmitting branch. These parameters are interrelated and conflicting. For example, the interdependence of the leakage power and the low-level loss may be computed on the basis of a somewhat idealized TR box as is done in Appendix A and the results presented in the form of the curves of Fig. 15. It is customary to design the cavities for matched input conditions ($\sigma = 1$), for obvious reasons, and for a low-level loss of one to two db. The relationship between the transmitting power dissipated in the TR tube and the low-level loss is shown
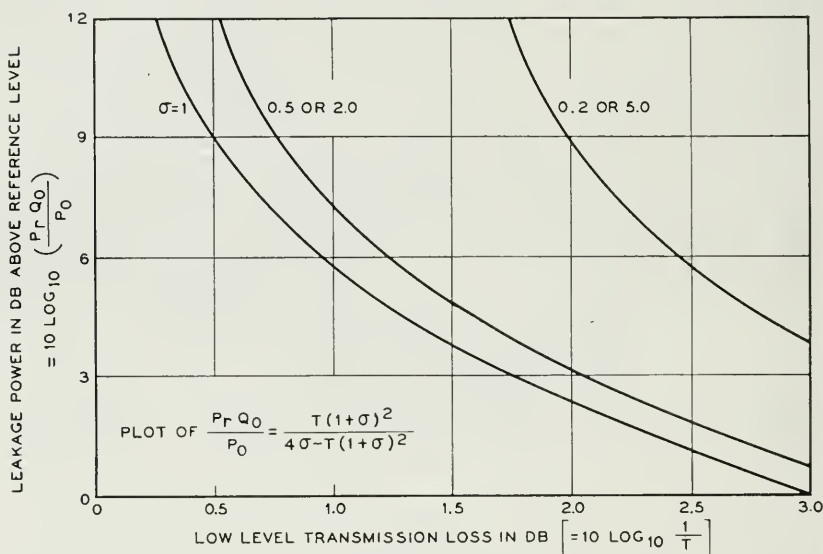
Fig. 15—The variation in leakage power with the low level loss adjustment

in Fig. 16. This curve may also be used to determine the effect of the low-level loss adjustment of a TR cavity on the recovery time characteristic since recovery time depends upon the gas discharge power rather than upon the transmitter power per se. In spite of this interdependence, it will be convenient to consider the different operating parameters separately in the sections to follow. The receiver protection aspect will be considered first.

Receiver Protection. The receiver protection problem is complicated by the fact that power reaches the receiver through the TR box by three different mechanisms. As shown in Fig. 17, the observed leakage power pattern is composed of three parts known respectively as the spike, the normal flat leakage and the direct coupling. The spike is the result of the transient

condition existing at the beginning of each pulse. Normal leakage power can be thought of as due to the finite voltage drop across the gas discharge while the direct coupling is that component of the leakage power which would be present if the voltage drop across the discharge were zero.

PLOT OF $P_g \left[\dfrac{Q_0}{PP_0} \right]^{\frac{1}{2}} = \left[\dfrac{4\sigma^2}{4\sigma - T(1+\sigma)^2} \right]^{\frac{1}{2}}$ FOR $\sigma = 1$

GAS POWER IS PROPORTIONAL TO SQUARE ROOT OF TRANSMITTER POWER

POWER DISSIPATED IN GAS IN DB ABOVE REFERENCE LEVEL

$= 10 \log_{10} P_g \left(\dfrac{Q_0}{PP_0} \right)^{\frac{1}{2}}$

LOW LEVEL TRANSMISSION LOSS IN DB $\left(= 10 \log_{10} \dfrac{1}{T}\right)$
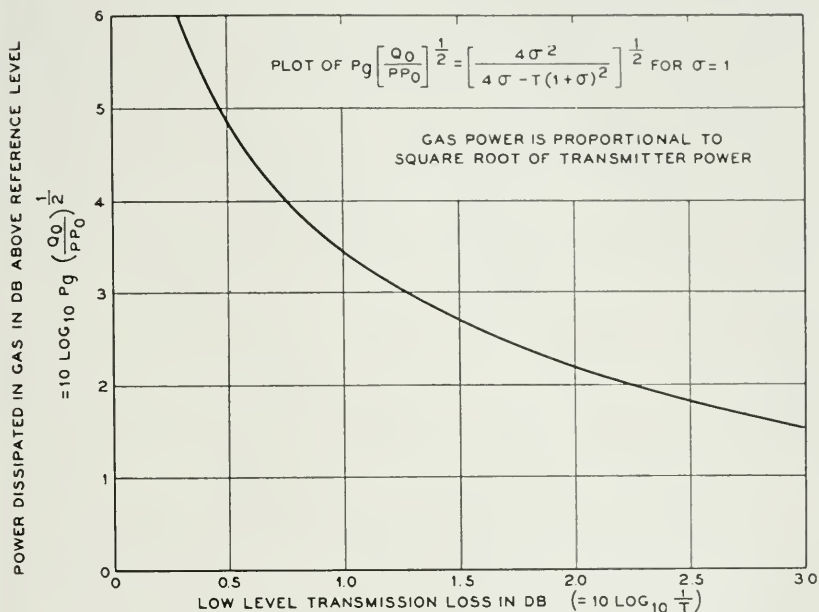
Fig. 16—The variation in gas discharge power with the low level loss adjustment

SPIKE

FLAT

PULSE AS DISPLAYED ON CRO WITH WIDE-BAND AMPLIFIER

SPIKE

FLAT

DIRECT COUPLING

PULSE LENGTH

IDEALIZED PULSE FOR COMPUTATION
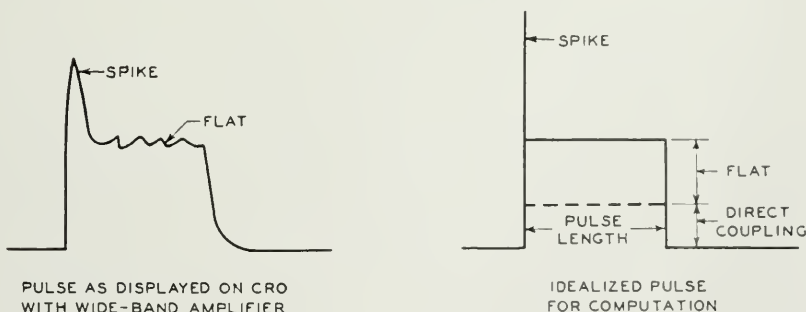
Fig. 17—The shape of the leakage power pulse

The Spike. Because of the rapid rate of rise and fall of the spike, the observable deflection on an oscilloscope is dependent upon the bandwidth of the video amplifier and upon the energy in the spike. Although an

observation of the true shape of the spike has not yet been made, the dura-
tion is estimated to be of the order of 10^{-9} seconds, a time interval that is
probably short compared with the thermal time constant of the contact on a
converter crystal.   However, it is possible to measure the energy content of
the spike, and such measurements indicate that this energy is fairly inde-
pendent of the length of the pulse and of the transmitted power level,
although it is definitely dependent upon the steepness of the wave front of
the transmitted pulse.   The spike clearly represents energy transfer through
the TR box during the period required to establish the discharge conditions
which exist during the flat.   The energy contained in the spike varies be-
tween a few hundredths of an erg to perhaps one erg per pulse, depending
upon a variety of factors.   By way of comparison, the conventional crystal
rectifiers are proof tested in manufacture with a single spike of 0.3 erg to
5.0 ergs, depending upon the crystal type.   It is generally believed that the
spike is more damaging than the flat in most radar systems.

The energy in the spike is found to depend upon the repetition rate of the
transmitting pulses, presumably because of residual ionization in the gas
discharge gap.   At low repetition rates (that is less than roughly 1,000
pulses per second), the spike energy may be materially decreased by a d-c
glow discharge near the radio frequency gap.   This discharge provides a
continuous supply of ions and free electrons and so aids in establishing the
desired condition in the r-f discharge path.   A discharge is supplied in all
the standard TR tubes.   An auxiliary electrode called the "igniter" or
"keep-alive" is used as the cathode, with the back or inside portion of one
of the high frequency electrodes acting as the anode.   A small amount of
radioactive material is placed in the tube to insure that the igniter discharge
starts on the application of the igniter voltage.   Fig. 18 is a plot of the way
in which the spike energy varies with the repetition rate both with and with-
out an igniter discharge.   Igniter oscillations sometimes occur as a result of
the negative resistance characteristics of the igniter discharge.   This causes
a cyclic variation in the number of free electrons and ions with a resulting
fluctuation in the spike energy.   Inadequate protection may result from
such oscillations.   It is customary to mount a current limiting resistance
very close to the igniter cap to minimize the effects of these undesirable
oscillations.   When such oscillations still occur they are usually evidence
of an insufficiently high igniter voltage or of tube failure.   The margin of
safety in the igniter operation may be increased by increasing the discharge
current but at the expense of greatly reduced tube life.

When a radar system is first turned on, the first pulse occurs without the
benefit of residual ions in the discharge, and for the first few pulses the spike
energy may easily reach dangerously high values.   While the magnitude
of this "turn on" effect is greatly reduced by the presence of the igniter

discharge, it is customary to provide a "crystal gate" in the form of a shutter which isolates the crystal from the TR box until after stable transmitting conditions have been reached and until the TR tube discharge has been established. The need for this additional turn-on protection is somewhat greater with the 724B than it is with the 721A tube. Another important function of the "crystal gate" is to prevent the crystal in an idle radar from being damaged by energy from other radars operating nearby.
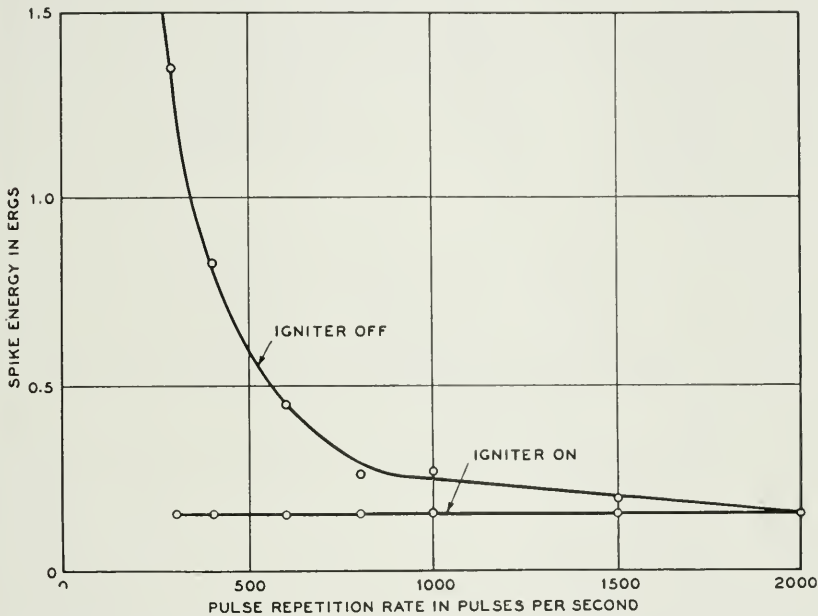
Fig. 18—The dependence of spike on the repetition rate for the 724B tube in a cavity adjusted for a 1.5 db low level loss and for a transmitter power level of 8 kw peak

The energy in the spike is a function of the effective size of the input and output coupling windows of the TR box. A convenient method of presenting this effect is to plot the spike energy as a function of the low-level transmission loss of the cavity which also depends upon the window sizes. Fig. 19 is such a plot for the 724B tube*. Comparing these experimental data with the computed flat power curve of Fig. 15, one notes that the spike energy varies at a more rapid rate than does the flat power. In both cases, the leakage decreases as the low-level loss increases and crystal protection can be purchased at the expense of receiver sensitivity.

* Based on data taken at the M.I.T. Radiation Laboratories by F. L. McMillan. Jr. and J. B. Wiesner.
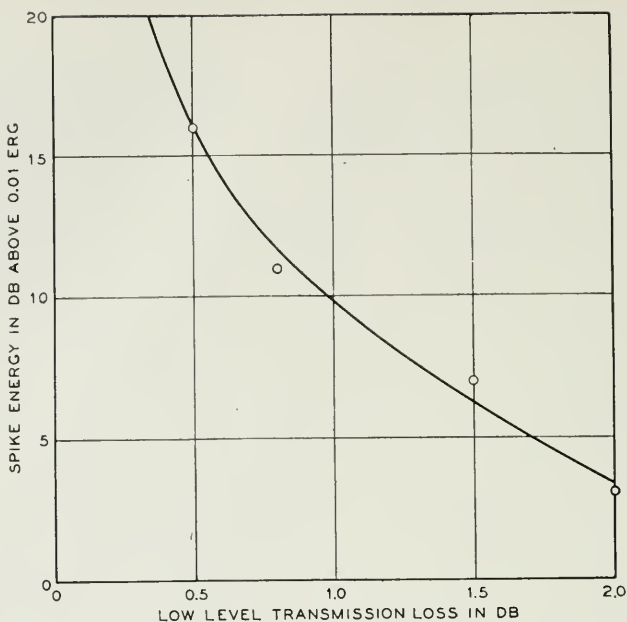
Fig. 19—The variation in spike energy with the low level loss adjustment ($\sigma = 1$) for the 724B. This experimental curve for the spike energy should be compared with the idealized flat power curves of Fig. 15
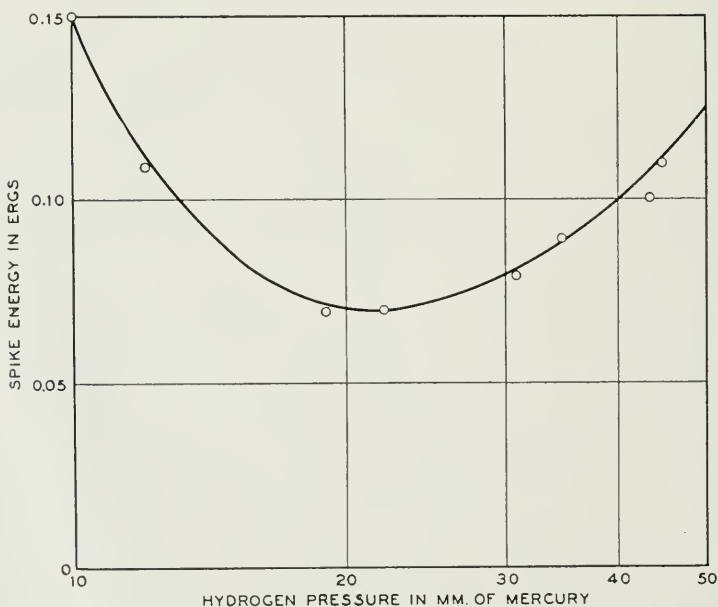
Fig. 20—The effect of gas pressure on the spike energy for the 724B

The way in which the spike energy varies with pressure of the gas in the TR tube is illustrated in Fig. 20. These data were obtained on the 724B tube structure. Other factors, yet to be discussed, prevent the use of the exact optimum pressure as determined on the basis of the spike energy only.

The Flat. The more or less flat portion of the leakage power is in reality the result of two different mechanisms of energy transfer, one of which is reasonably independent of the transmitter power level. It is this portion only with which we will now be concerned. This flat power is critically dependent on the chemical constitution and pressure of the gas within the TR tube. It can be thought of as being the power transmitted by the TR
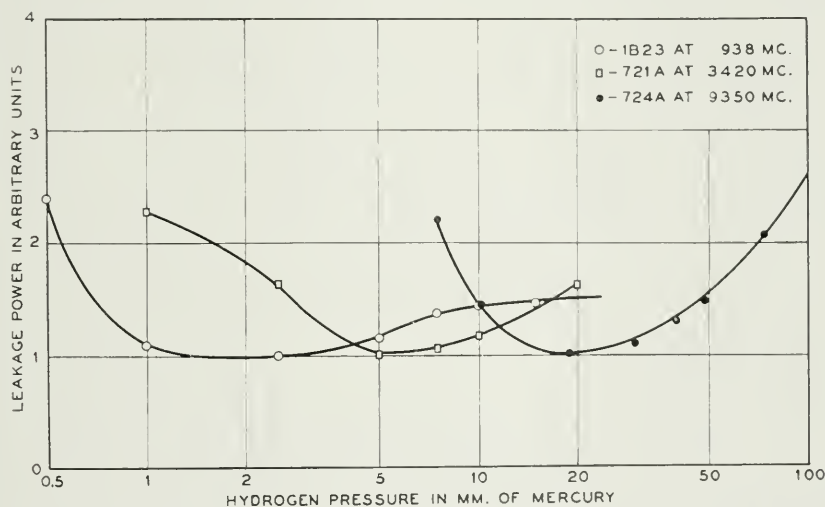
Fig. 21—Experimental curves showing the relationship between flat leakage power and gas pressure, taken with a c-w oscillator

box by virtue of the fact that the voltage drop across the gas discharge is not zero. The constancy of the flat power in spite of variations in the transmitter power level is presumably related to the similar phenomenon of a nearly constant voltage drop across a d-c gas discharge independent of the discharge current. Because of this constancy, the gas discharge parameter P_0, shown in Fig. 15, can be assumed to be a constant more-or-less independent of the transmitter power level. Reasonable values of P_0 for cavity design purposes are 20 volt-amperes for the 721A tube and 10 volt-amperes for the 724B tube. Corresponding values of the Q_0 parameters needed in interpreting Fig. 15 are 2500 for the 721A tube and 1500 for the 724B tube. Using these values the flat leakage power for a TR box using a 721A tube

and having a low-level loss of 1 db would be 30 milliwatts. The corresponding flat leakage power for the 724B tube in a 1.5 db box would be 16 milliwatts. Actual measured values are usually somewhat less than these figures. As most crystals will withstand flat powers very much greater than this amount, the flat power is normally of much less importance than the spike in contributing to converter crystal failure.

Since the flat portion of the leakage power represents quasi-steady-state conditions, it is possible to simulate it for purposes of study by the use of a C.W. oscillator. Fig. 21 contains three experimental curves taken at
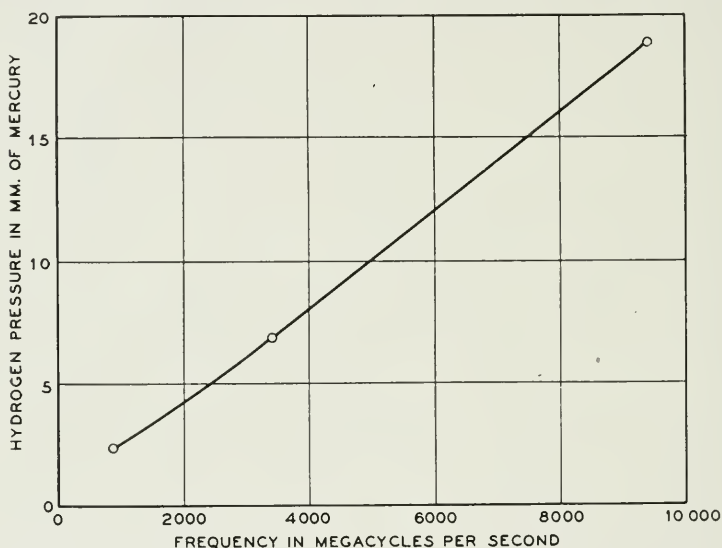
Fig. 22—The pressure for minimum leakage power as a function of frequency

three different frequencies showing the relationship between the flat leakage power and gas pressure. These curves were all taken with tubes filled with hydrogen only. Fig. 22 shows that the pressure for minimum flat leakage is proportional to the frequency. This simple law probably does not apply at frequencies much less than 1000 mc.

Water vapor is used in commercial TR tubes to improve the recovery time, as will be discussed later. The variation in flat leakage power with partial water vapor pressure as measured on a 721A type of tube containing both hydrogen and water is shown in Fig. 23. These data were taken in a radar system.

In this connection, it is of interest to note that the characteristics of the gas discharge in the TR box must of necessity be quite different from those that obtain at lower frequencies. Simple calculations indicate that the

mean free path of an electron is in general of the same order as the distance
between the electrodes but that very few electrons are able to reach the
electrodes because of the very rapid reversals in the r-f field. Electrons
therefore oscillate rapidly to and fro, losing energy to the neutral gas mole-
cules and to positive ions through occasional collisions The positive ions
do not contribute in any substantial way to the discharge current because
of their large mass and correspondingly low velocity. The r-f voltage drop
across the discharge is maintained at a relatively low value by the formation
of more ions and free electrons by collisions between electrons and neutral
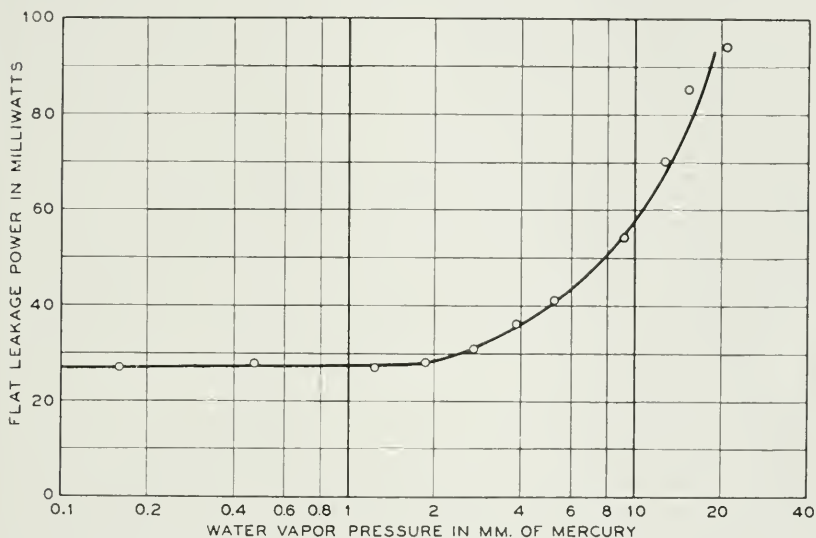
Fig. 23—The effect on leakage power of the addition of water vapor to 20 mm of hydrogen
in the 721A type tube

molecules as soon as this voltage rises above some critical value. Measure-
ments indicate that the voltage drop across the r-f gap is of the order of 80
to 100 volts for a typical TR tube. The variation in voltage drop with gap
length may be inferred from the flat power measurements recorded in Fig.
24.

Direct Coupling. At very high transmitter power levels a third component
of leakage power is observed which is directly proportional to the trans-
mitter power. This component is usually called "direct coupling". It is
due to the transmission of power through the cavity in modes which do not
have voltage maxima at the gas discharge gap. It can therefore be ob-
served even when the gap in the tube is short circuited. In fact measure-
ments made under such short-circuited gap conditions yield results com-

parable to the values observed for actual tubes.   The direct coupling component of the total flat power and the gas discharge limited component are found to be additive.   Direct coupling power is logically measured in terms of db below the transmitter power level and for the usual TR box is of the order of 60 to 70 db.   An abnormal form of direct coupling which may reach dangerous values can occur under certain improper operating conditions when the magnetron produces an appreciable amount of power at other than the normal operating frequency.   Some of these spurious frequencies may be in the vicinity of the resonant frequencies for these "direct coupling"
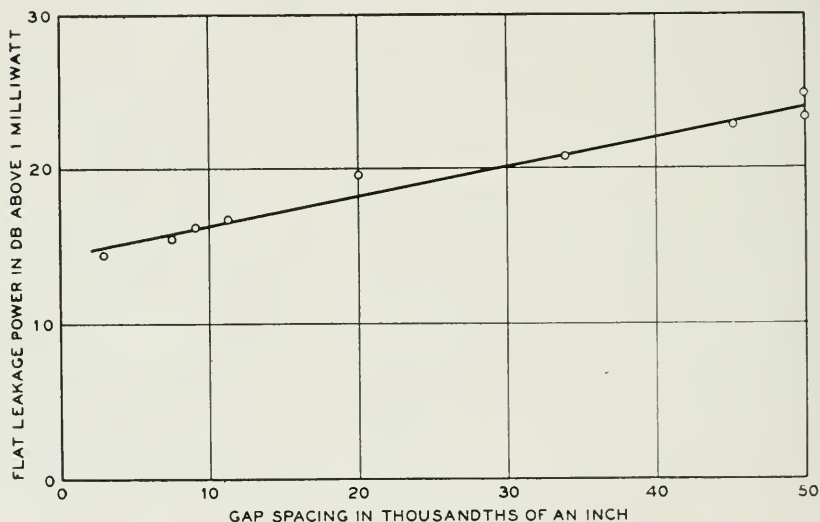
Fig. 24—The variation in flat leakage power with gap length, measured at 3000 megacycles

modes in the TR box and so be transmitted without much attenuation. Normally, direct coupling is of interest only in very high power systems.

Receiver Self-Protection.   The fact was mentioned earlier that a receiver can provide itself with a certain amount of self-protection as a result of its change of impedance with level.   This effect is still of use even in systems employing TR boxes.   Unfortunately the apparent source impedance at the TR box output terminals is different for the different components of the leakage power so that the self-protection feature cannot be utilized for all components simultaneously.   The matter is further complicated by the fact that the converter crystals themselves vary greatly in their impedance and in their variation of impedance with power level.   The best designed converters as far as crystal protection is concerned are usually those which

provide a certain amount of self-protection against the spike. It has been estimated that this self-protection seldom exceeds 2 db in practice.

Leakage Power Measurement. The c-w method of measuring the flat power has already been mentioned. Spike energy and direct coupling must, of course, be measured under normal high level operating conditions. Relative measurements of the spike can be made with an oscilloscope, acting ballistically, and the factors which affect the spikes can be studied in this manner. A correlation between the relative spike energies and the degree of crystal protection can be obtained by trial and from this correlation the operating conditions for adequate protection can be determined. Most of the early studies were made in this way. It is possible to deduce absolute values for spike energy, flat power, and direct coupling from measurements made when all three are present because of the different ways in which these parameters vary with the recurrence rate, pulse length and transmitter power. The method of doing this is outlined in appendix D.

A more precise method of measuring the spike energy involves the cancellation of the flat power by a signal of adjustable phase and amplitude obtained from the high-level transmission line. The average spike power is then measured directly and energy per spike computed. Most of the spike data quoted earlier were obtained in this fashion.

High-Level Loss. The power dissipated in the TR box as a result of the gas discharge is not ordinarily a large enough fraction of the total transmitter power to be of any concern. Using the P_0 values previously quoted, it is possible to compute the gas discharge power by the use of Fig. 16. At a line power of 100 kw and a low-level loss of 1 db the gas power in the 721A tube is 63 watts. The corresponding figure for a 1.5 db box using the 724B tube is 47 watts. For these cases the high-level loss is therefore less than 0.005 db. Low as this fraction is in db it still may be high enough to affect the life of the TR tube, as discussed in a later section. No trouble of this sort is ordinarily encountered with the 724B or 721A tubes. The chief cause of failure of the 1B23 is from loss of Q and this in turn is caused by the sputtering action of the high-frequency discharge.

Recovery Time. As mentioned earlier a TR box must recover its low-level properties at the end of the transmitting pulse in a very short period of time. The actual "recovery time" is in fact several orders of magnitude smaller than the deionization times of the usual gas discharge so that a quite different mechanism must be involved. While an exact theory of the recovery is beyond the scope of the present paper, a qualitative picture of the recovery process may be of interest.

During the transmitting period the free electrons provide almost all of the discharge current, and are replenished by electron-molecule collisions. At

the end of the transmitting period these electrons may migrate from the discharge region, they may recombine with the positive ions, or they may be captured by molecules to form negative ions. Negative ion formation by attachment effectively removes an electron from the discharge because of the great increase in mass. It is an experimental fact that those gases which readily form such ions (of which water vapor is the most common) are the gases which exhibit good recovery in a TR box. This process is
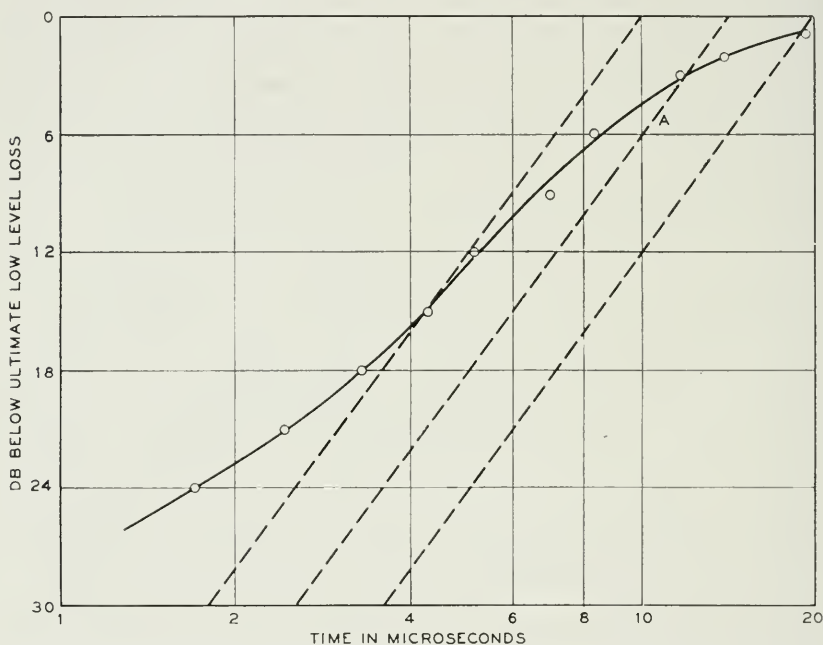
Fig. 25—A typical recovery time characteristic for the 721A tube in a TR cavity adjusted to 1.5 db low level loss with a transmitter power level of 100 kw peak

not deionization in the ordinary sense and it can take place at a surprisingly rapid rate.

Of course, immediately upon the termination of the transmitting pulse, the cloud of free electrons will cause an extremely high loss to any reflected signal but the loss will rapidly decrease to some limiting value set by the fixed losses in the TR cavity itself.

A typical recovery curve for the 721A tube is shown in Fig. 25. This curve has a particularly fortunate shape in that the variation in loss with distance, or more correctly with time, is at approximately the same rate as the variation in the reflected signal level with distance for a target of fixed size. The importance of this can be understood by considering the way in

which the reflected signal intensity varies as an object of fixed size approaches a radar set from a great distance. Such an object as seen by a given radar set may be represented on Fig. 25 by a straight line having a slope of 12 db per factor of two in distance. Several such lines are shown. Considering line A it will be observed that this target can first be seen at a distance corresponding to 12 microseconds. Since this target line always remains below the TR recovery line for times shorter than that at the intersection point, an object once seen will remain in view continuously as it approaches
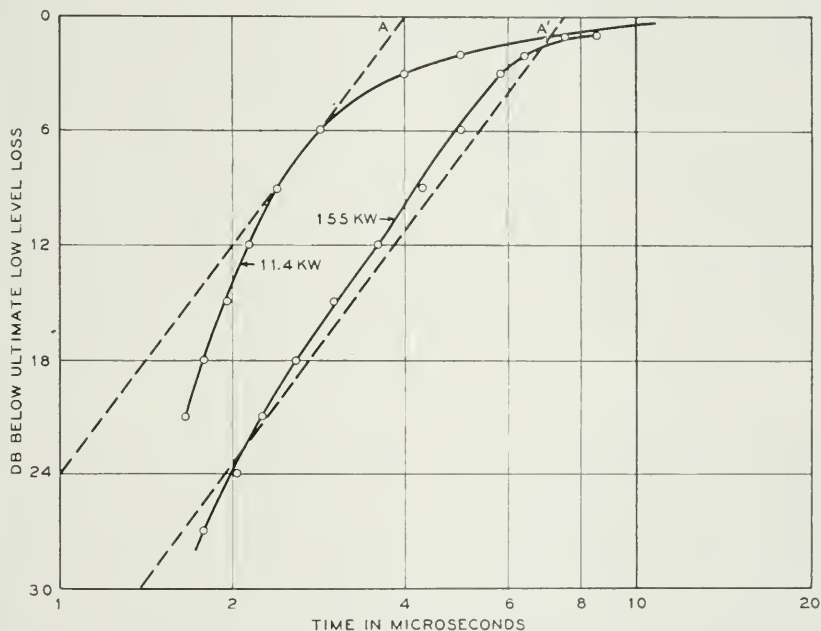
Fig. 26—The recovery time characteristic at two different transmitter power levels for the 724B tube in a 1.5 db TR box

the transmitter even in spite of the poor recovery characteristics of the TR box. A certain amount of sensitivity-time-control action is actually provided. The recovery time characteristic frequently does not necessarily set a lower limit on the effective range of a radar system although it always sets a limit on the smallest effective target size which can be observed.

The recovery time characteristic is critically related to the transmitter output power level as shown by the data for a 724B tube shown in Fig. 26. When the recovery time curves for two different power levels are compared, the target line which is just detectable at a power level of 11.4 kw is shown as the line marked A   Contrasting with the 721A behavior this target would be visible only at one point and would then be lost from sight as it approached

the transmitter. This same target is represented on the plot as line A'
for a transmitter output power of 155 kw, that is being displaced vertically
by approximately 12 db to take account of the difference in transmitter
power. At this higher power level the target would be visible at a much
greater distance (corresponding to 7 microseconds elapsed time) and would
remain in view until the target distance corresponded to 2.1 microseconds
time. In this case an increase in power by 12 db resulted in an increase
in range by a factor of 2.8. While this would indicate that an increase in
range can be obtained by increasing the transmitter power, it should not be
inferred that an increase in the near-range sensitivity will always result from
an increase in power. At any specified range there appears to be a unique
value of transmitter power output beyond which the loss in TR box recovery
more than offsets any increase in range due to higher output powers. While
accurate figures are not available for the 721A tube, there is some evidence
that an output power of 100 kw is already too large for ranges corresponding
to elapsed times of 10 microseconds or less. Under these conditions im-
proved operation results from a decrease in the transmitter power level.
Such an effect has never been observed by the writers with the 724B tube,
probably because the transmitter powers available in its operating fre-
quency range have usually been somewhat less than that available with the
721A tube.

It should be noted, at this point, that the recovery time does not depend
upon the transmitter power only, but rather upon the gas discharge power
which is a function of both the transmitter power and the low-level loss
adjustment of the TR box as shown by Fig. 16. A very great improvement
in near range sensitivity can usually be obtained by increasing the trans-
mitter power level and at the same time increasing the low-level loss adjust-
ment of the cavity to limit the gas power to a value for which the recovery
time is satisfactory. This of course increases the ultimate low-level loss and
so adversely affects the long-range sensitivity.

The dependence of the recovery time on the ambient temperature for the
721A tube is shown in Fig. 27. The 724B tube is much less temperature
dependent. This variation in recovery time with temperature is caused
by the reduction in water vapor pressure through condensation, as shown
by the identity of the recovery curve for a standard 721A tube at $-186°$ C.
with a special tube filled with hydrogen only.

With continued life the water vapor content of the tube decreases with
the corresponding change in the recovery time characteristic. Fig. 28
shows the effect with the 721A tube. The dependence of the recovery time
on the water vapor content in the 724B tube is shown in Fig. 29. Comparing
this curve with Fig. 27, it will be observed that the loss of water vapor has
much less effect on the recovery characteristics of the 724B tube than on the

721A tube.   It should be noted, however, that the 724B tube frequently reaches the end of its useful life as a result of its failure to provide adequate receiver protection before serious loss of recovery occurs.

The ATR, if one is used, can also contribute to poor recovery as may be seen by referring to Fig. 30.   These data are not necessarily representative since it is possible to adjust the length of line between the magnetron
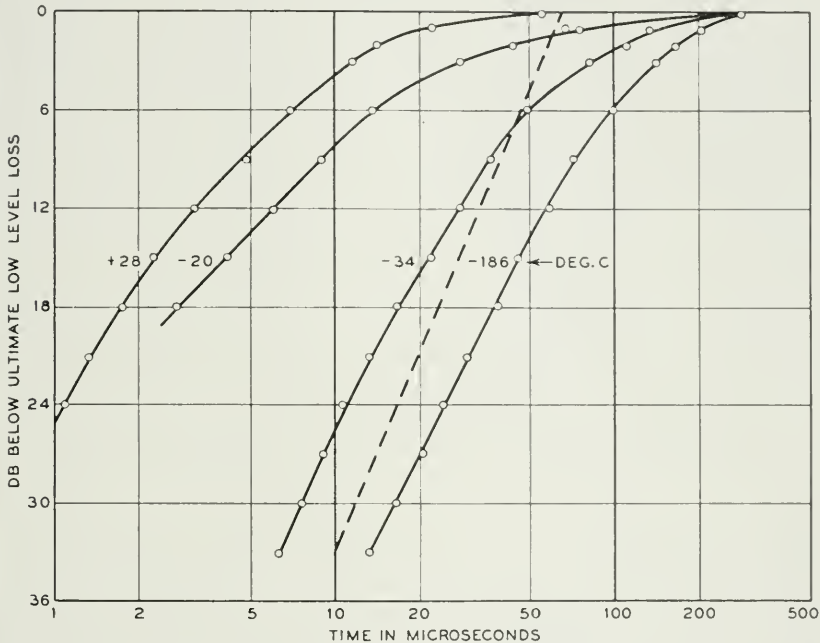
Fig. 27—The dependence of the recovery time on the ambient temperature for the 721A tube

and the ATR junction so as to minimize the effect.   Nevertheless the effect is important and should not be overlooked.

Low-Level Loss.   An analysis of the low-level loss must take into account two components of loss, the first resulting from power loss in the TR cavity itself and the second resulting from the fact that some power will always be absorbed by the transmitting branch of the system.

The relationships existing between the low-level loss adjustment of a TR box and its other performance characteristics have already been discussed. One aspect of the problem, not previously considered, has to do with the dependence of the performance on the Q of the cavity.   This is clearly shown in Fig. 15 which has already been referred to in a different connection.   From this aspect, at least, the higher the Q of the tube and its associated cavity the
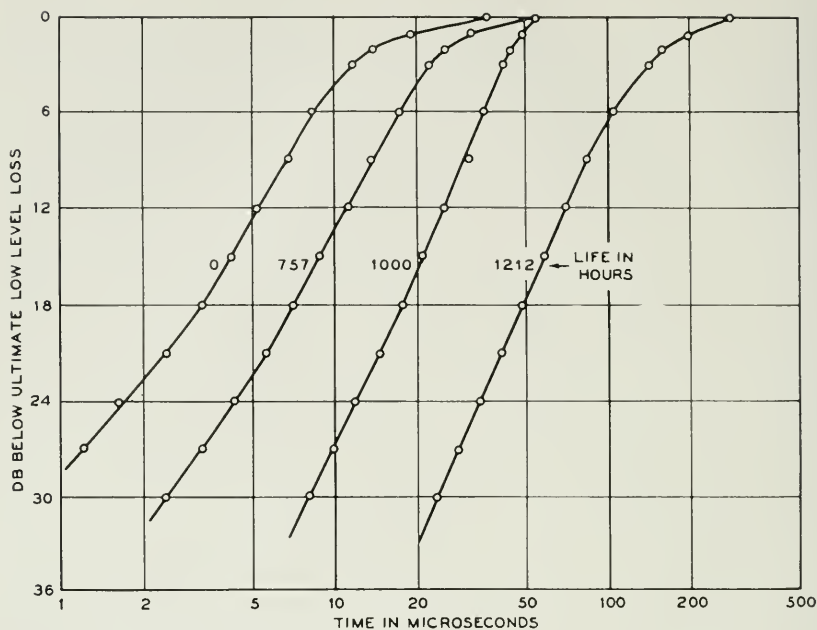
Fig. 28—The variation in recovery time with life for a typical 721A tube
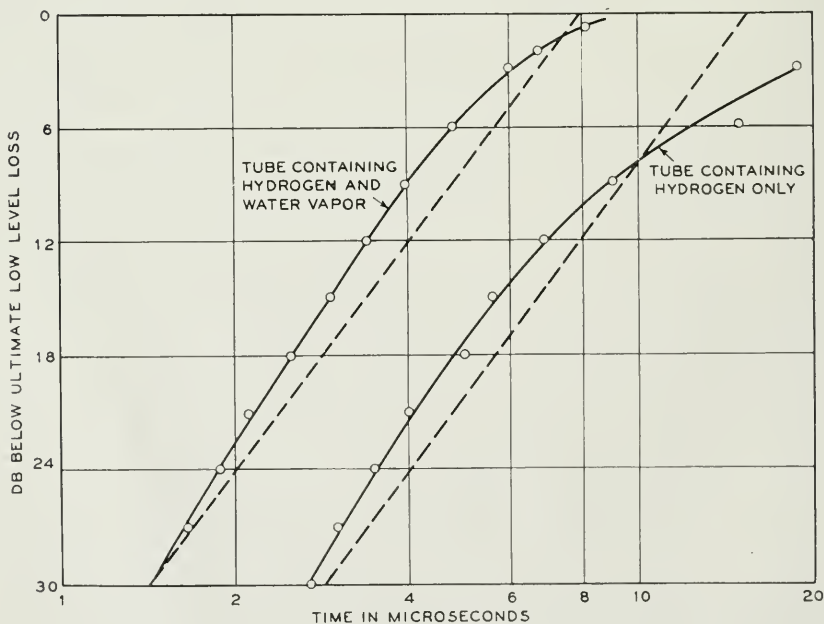
Fig. 29—The variation in recovery time with gas content for the 724B tube

better the over-all performance of the system.* Any basic improvement in
Q can be reflected either in a lowering of the leakage power or in a reduction
of the low-level loss, as may be desired.

Variations in the Q between tubes used in a given cavity of fixed design,
are, however, seen as variations in the low-level loss only and have no
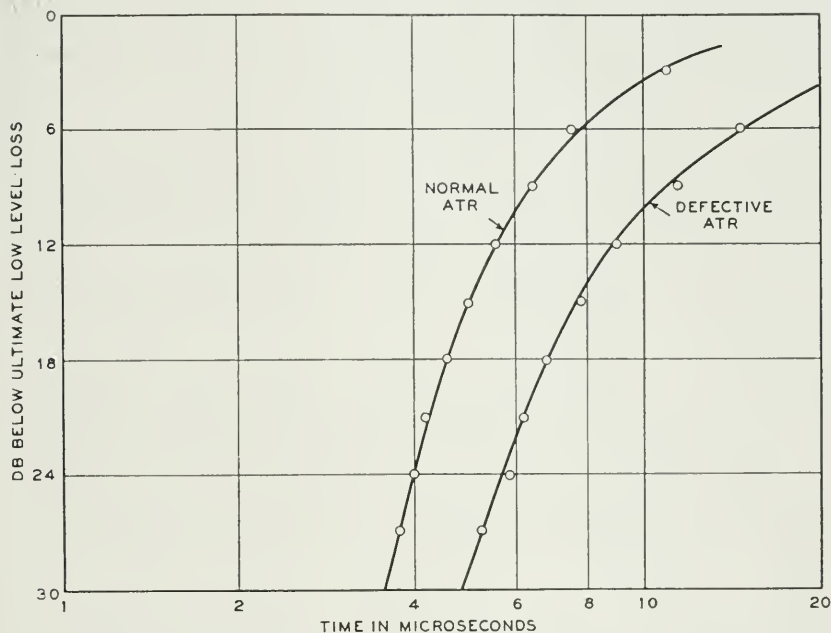
Fig. 30—The ATR recovery effect

noticeable effect on the leakage power. This may be understood by refer-
ence to the equations derived in appendix A. The performance of a some-
what idealized TR box may be expressed in terms of three design parameters
δ_0, δ_1 and δ_2 which relate respectively to the properties of the cavity, its
input coupling, and its output coupling; and in terms of a gas discharge
parameter P_0. In terms of these new parameters the in-tune low-level
tansmission of a TR box is given as a power ratio by

$$T = \frac{4\delta_1\delta_2}{[\delta_0 + \delta_1 + \delta_2]^2}. \qquad (12)†$$

* As explained later in this section, band width limitations set an upper limit to the
permissible Q.
† Numbered equations in the text correspond with the numbers used in the appendices.

The input standing wave ratio on the input line is given by

$$\sigma = \frac{\delta_1}{\delta_0 + \delta_2}. \tag{13}$$

The leakage power is similarly given by

$$P_r = P_0 \, \delta_2 \tag{14}$$

while the gas-discharge power in terms of these parameters and the power in the transmitting line (P) is given by

$$P_g \doteq (P \, P_0 \, \delta_1)^{1/2} \tag{19}$$

The parameters δ_1 and δ_2 are properties of the input and output coupling as they are geometrically related to the cavity and are substantially independent of the Q of the cavity. The parameter δ_0 is, however, the reciprocal of the intrinsic or unloaded cavity Q. Equation 12 is seen to depend upon δ_0 but equations (14) and (19) do not. The effect of variations in Q is thus demonstrated.

The over-all performance is also affected by the relative values of δ_1 and δ_2. In view of the dependence of P_r on δ_2 directly and on δ_1 indirectly through the fact that P_0 is not entirely independent of P_g, it is advantageous to adjust the values of δ_1 and δ_2 so that the input standing wave ratio (σ of equation 13) is unity. Such a condition is also very desirable for system reasons as well. When this condition is met, equation (12) reduces to

$$T = \frac{\delta_2}{\delta_1}.$$

The curve marked $\sigma = 1$ of Fig. 15 and the curve of Fig. 16 were plotted on this basis. It should be noted that matched input requires that the input window be larger than the output window.

TR boxes are unfortunately not always operated in the in-tune condition, and they must also pass a band of frequencies as fixed by the narrowness of the transmitter pulse. For these reasons the Q must not be set at too high a value. The additional low-level loss which results from off-tune operation may be computed from equation 28 of appendix A.

Incidentally, it is an experimental fact that the leakage power and the gas-discharge power are not materially altered by small departures from the in-tune adjustment, presumably because of the very low effective Q of the gas discharge.

The ATR Low-Level Loss Component. The component of low-level loss which results from losses of power to the transmitting branch depends very greatly upon the "cold impedance" of the magnetron or other transmitting tube and upon the properties of an ATR box if one is used. As shown in

appendix C the loss chargeable to the ATR and the associated transmitting arm can be expressed as a factor F given by

$$F = \frac{4}{(2 + G)^2 + B^2} \tag{33}$$

where G and B are respectively the conductance and susceptance of this branch in units of the surge admittance of the transmission line. Since
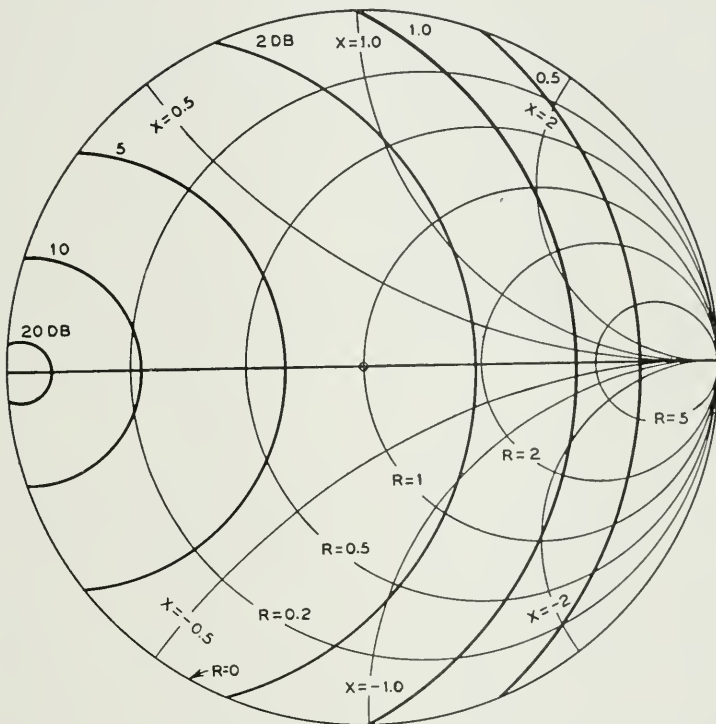
Fig. 31—Curves for constant ATR loss in db as a function of the impedance of the transmitting branch

curves for constant values of F appear on the reflection coefficient plane (Smith transmission line chart) as circles this presentation is very convenient. Fig. 31 is such a plot (impedance circles rather than admittance circles are shown).

If now an ATR is introduced having a resistive component of impedance the range of values of G and B is restricted so that a minimum value of F exists for any random value of the magnetron impedance. With variations in the magnetron cold impedance or in the effective length of line between the magnetron and the ATR junction the value of F will vary between this

minimum value and some maximum value which may approach unity. For example if the ATR is adjusted to have the same gas-discharge power as that in a TR adjusted for a transmission of T, its low-level in-tune input impedance will be

$$Z = \frac{1}{1 - T}. \tag{38}$$
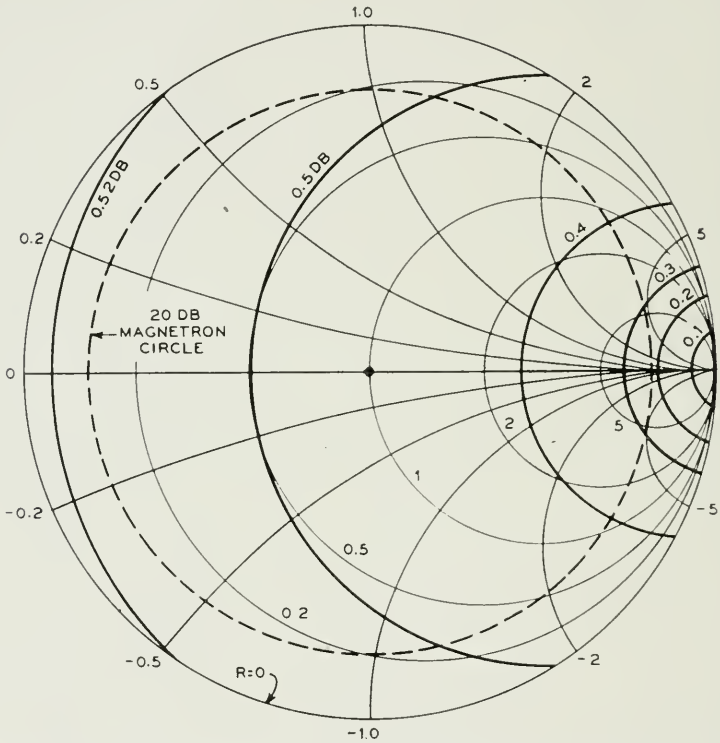
Fig. 32—The ATR low level loss as a function of the magnetron impedance with an ATR adjusted to an impedance of 8 + j0

Actually since the gas-discharge power is usually not the limiting factor in the design it is possible to adjust the ATR to have an input impedance of 8 or 9 (in terms of the line characteristic impedance), corresponding to a TR low-level transmission of the order of 1/2 db ($T = 0.89$) and yielding an ATR loss of approximately 1/2 db.

Since the exact value of the impedance of the magnetron branch is not necessarily known it is convenient to show the dependence of the loss factor F for any given ATR on the magnetron impedance by a plot somewhat

similar to Fig. 31 but transformed to the magnetron side of the ATR junc-
tion. This may be done by subtracting the ATR impedance from the values
read off of Fig. 31 corresponding to desired values of F and replotting these
on the reflection coefficient plane. As an example, the in-tune value of Z
for one typical 724B ATR cavity is $8 + j0$. Points lying on the $R = 8$
circle on Fig. 31 will then lie on the magnetron $R = 0$ circle, the region
inside being distorted and expanded to fill the entire positive R region on
the reflection plane. The results are shown in Fig. 32. From this plot it
is evident that the maximum possible low-level loss chargeable to the trans-
mitting branch would be slightly more than 0.52 db and that this would
occur only for a restricted range in the value of magnetron impedance. As a
matter of practical interest the "cold impedance" of the usual magnetron
is such as to give at most a 20-db standing wave. This restricts the possible
range in impedance values to the area on Fig. 32 within the dotted circle,
thus limiting the maximum loss to slightly less than 0.52 db, and imposing
a minimum loss limit of 0.22 db.

This type of analysis may be extended to consider the ATR loss during
the recovery period if desired although the problem becomes rather com-
plicated as a result of the simultaneous variation in input impedance of both
the TR and the ATR.

TR Box Design Considerations

The desired electrical properties for a TR box can of course be achieved in
a variety of different physical structures. A construction technique which
separates the gas-discharge tube from the rest of the TR box cavity offers
many advantages. In the first place the cost of the entire device is kept low
by reason of the fact that it is not necessary to transmit the tuning motion
through the vacuum-tight tube enclosure. The replacement cost is also
greatly reduced since the more complicated part of the TR box is a perma-
nent part of the equipment. Then, the same tube structure can be used
for a variety of different types of equipment operating in different wave-
length bands and requiring different amounts of receiver protection by the
use of different size cavities and different size coupling windows. This
greatly simplifies the problem of maintaining replacement stocks. An addi-
tional factor, which was of importance during the early days of the war,
is that the design of such a tube can be frozen at an early stage, before all
the possible circuit aspects of the TR problem have been solved since changes
in the external parts of the TR box can be made independent of the design
of the replaceable tube element. The widespread use of the 721A and 724B
vacuum tube is, in a sense, proof of the essential soundness of the arguments
for the external cavity type of construction.

The chief difficulty to be overcome in the design of a separate cavity type of TR box has to do with the need for a low-loss contact between the internal portions of the tube and the external cavity. A copper-disc sealing technique, developed at the Bell Laboratories in connection with the construction of water cooled tubes* and later superseded by the now conventional Housekeeper seal, had previously been applied at ultra-high frequencies in the design of oscillators and amplifiers. This technique makes possible very satisfactory high-frequency connections by simply clamping the external portion of the disc between machined surfaces. The flexibility of the copper discs is sufficient to compensate for minor machining errors and for differential thermal expansions while the relative softness of the copper insures a continuous contact around the entire periphery. The goodness of contact provided by these contacts is evidenced by the fact that Q's of 4000 and greater are obtained at 3000 megacycles with discs of the 721A type. This technique was therefore adopted for the 721A tube and the 724B tube and for one electrode of the 1B23 tube. The second high-frequency electrode of the 1B23 was made in the form of a rod terminating in a ball for convenience in replacing tubes since the accompanying loss of Q can be tolerated in the frequency range where this tube is used.

In an external cavity type of TR box the over-all goodness of the design is largely determined by the design of the gas-discharge tube. It is the tube designer's responsibility to determine the optimum shape and size for the copper discs and for the glass tube envelope and to determine the optimum gas composition and pressure, with due consideration being given to such matters as mechanical ruggedness, manufactureability and freedom from undesirable ambient temperature, pressure and humidity effects.

With the copper-disc type of tube the system designer has at his disposal the ability to vary the design of the external cavity, and to arrive at any specific compromise between the various conflicting performance criteria which he feels to be the best for his particular application. For example, in systems employing vacuum tube converters it is customary to adjust the TR box for a low-level loss of 1 db or somewhat less since receiver protection is of minor interest while in systems employing crystal converters it is customary to fix the low-level loss at 1.5 db or sometimes as high as 2.0 db. Certain cavities, notably the one shown in Fig. 5, have to be designed to have an extended tuning range, in this case achieved by a piston tuner with, however, some loss in Q, while other cavities, the one shown in Fig. 11 being typical, do not require this same tuning range and a different tuning mechanism (in this case, tuning plugs) can be employed.

An extreme example, illustrating the advantages to the system designer

---

* W. Wilson, "A New Type of High-Power Tube," *B. S. T. J.*, vol. 1, p. 4, July 1922.

of the external cavity type of tube, is that of certain radar systems which were required to be capable of receiving signals on occasion at a frequency differing from their normal tuning. This was done by a solenoid-operated plunger which could be preset to alter the tuning of the cavity by the desired amount whenever the solenoid was energized.

THE TUBE DESIGN

The 702A and 709A vacuum tubes, as previously mentioned, were put into service with little or no consideration of their real suitability. With these stop-gap designs in production the basic design problem was given serious consideration, with separate studies being made of the mechanical design considerations as they relate to the size and shape of the discs and glass of the tube, and of the gas filling.

The exact shape of the disc is determined first by the total tuning range which is to be required of the tube, and second by the necessity for maintaining the Q of the structure as high as possible. It has been shown that in a spherical resonator with coaxial cones the maximum Q occurs when the cone half-angle is nine degrees. The copper-disc tube can only roughly approximate the ideal spherical resonator; nevertheless it appears desirable to use cones of this angle. The disc spacings and diameters are so chosen that the tube resonates at the shortest wavelength at which it is to be used in a "square" cavity; i.e., one in which the inside diameter approximately equals the height. Such a cavity is about the closest practical approach to a sphere. The glass diameter is made as large as mechanical considerations permit so it is as far as possible removed from the region of high electric field intensity.

The experimental results of Fig. 24, previously noted, indicate that the leakage power of a TR box decreases as the gap spacing decreases; thus one is tempted to make the gap extremely small. Too small a gap is very troublesome, however, since such a gap has an unreasonably rapid variation of resonant frequency with gap separation, making the tuning extremely subject to change as a result of dimensional variations due to processing or to temperature changes. Accordingly one chooses a compromise gap separation. The electrode radius at the gap must be large enough to permit the radio frequency glow discharge to dissipate the required power without excessive spreading, and must be determined by experiment.

Rather than attempting to hold all of the mechanical variations in the tube (including glass thickness) to the necessary tolerances to insure the desired uniformity in tuning, the tubes are pretuned before exhaust by deforming the copper discs. The tubes are placed in a special cavity and the disc inside the envelope distorted by a tool until resonance is obtained at a speci-

fied frequency. It is quite easy to tune tubes in this way so that they are uniform to within ± 0.25%.

Unless the tube is properly designed, changes in ambient temperature may seriously affect its resonant frequency. The part of the disc which is inside the glass envelope may be considered as a diaphragm supported around its periphery by the glass which has a temperature coefficient of expansion negligibly small compared to that of the copper. An increase in temperature, which causes the copper to expand, will force the cone tip to move toward or away from the gap, depending on the initial slope of the nearly flat portion of the disc. The temperature coefficient of frequency may be either positive or negative, and will have extreme variations in magnitude from tube to tube if consideration is not given in disc design to avoid such difficulties.

A cavity made wholly of copper will have a fractional change in wavelength with temperature the same as the fractional change in length of copper (approximately fourteen parts in a million per degree centigrade). As the temperature increases, the frequency decreases. At a frequency of 1000 mc, the approximate temperature coefficient of frequency is − .014 megacycles per degree centigrade; at 3000 mc it is − .042 mc/°C; and at 10,000 mc it is − .14 mc/°C. Magnetrons normally have temperature coefficients of about these magnitudes. The ideal TR tube would have the same coefficient as the magnetron; practically speaking, any coefficient between zero and twice the value for copper is satisfactory.

It is practical to make a copper disc structure which has the required temperature coefficient. Fig. 33, which is a cross-section of a 721A tube, shows how temperature compensation within the tube is effected. The disc is slanted away from the center portion of the tube, so that as temperature rises the cone is carried away from the gap. At the same time the cone itself expands; the net effect is to increase the gap between the two cone tips. The angle of the slanted part of the disc must be such that the gap increases with temperature at the same rate that it would in an all-copper cavity. If this condition is fulfilled, the net result of the expansion of all the tube parts, and of the cavity itself, will be the same as if it were all made of copper. This result is achieved by an experimental series of successive approximations. A number of models are built until the angles are found which give the desired temperature coefficient of frequency.

The gas content of the tube was the subject of considerable study. As stated in the section on Recovery Time, gases which readily form negative ions are invariably the most satisfactory from that viewpoint. Gases of low ionization voltage, such as the rare gases, give excellent protection but usually have extremely poor recovery. The choice of a TR gas must of necessity be a compromise between the two requirements. Some otherwise

satisfactory gases are not useful because of other characteristics. HCl is an excellent TR gas, but is very corrosive. Freon, a common refrigerant, is excellent but is unstable. In general, no gas which contains a solid elementary constituent is a satisfactory TR gas. The most satisfactory gas found was water vapor. It is cheap, stable, and easy to handle. Water vapor alone is not safe to use at a low temperature, so a small amount of hydrogen is added to ensure adequate protection when the water vapor is frozen out.

POSITION OF CONE AT
ELEVATED TEMPERATURE
(EXAGGERATED)
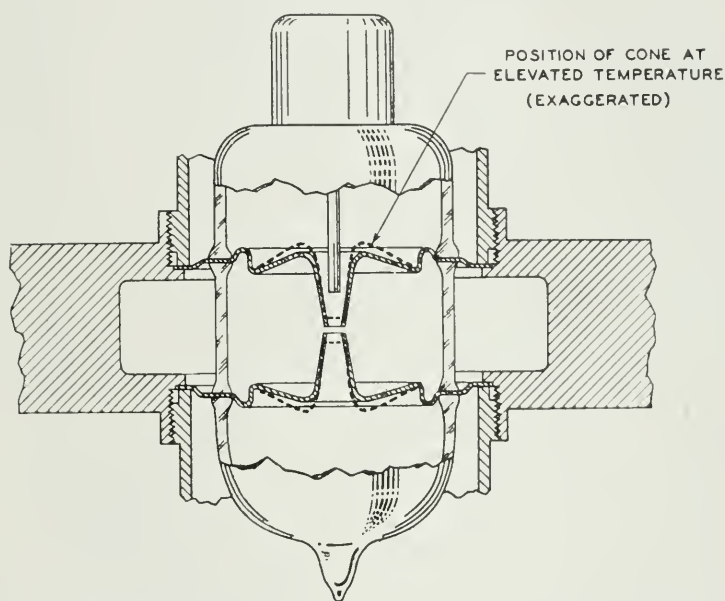
Fig. 33—Cross section of a 721A TR Tube, showing the special shape of the temperature compensated discs

The life of a TR box is in general determined by the gas volume. The radio-frequency discharge consumes no gas except at extremely high power levels; the igniter discharge accounts for the greater part of the loss of the gas initially placed in the tube. Reduction of the water vapor to hydrogen by formation of copper oxide on copper parts of the tube seems to be the principal process which goes on. This change results in no change in total pressure until the water vapor is exhausted; thereafter sputtering becomes more important and accounts for a fairly rapid hydrogen clean-up. The life of a TR tube is determined by the igniter current, which is maintained at a value as small as possible consistent with adequate spike protection, and by the volume of the tube.

Fig. 34—Setup for making 724B copper-to-glass disc seal

The lack of water vapor in a tube which has been operated for some hundreds of hours may manifest itself by a failure in either protection or recovery time. The operating frequency determines which failure becomes important first; at long wavelength it is likely to be recovery, while at shorter wavelength the spike protection is likely to fail first.

The life of a TR tube operated without igniter is very much longer. This may be understood from the picture given above under "recovery," of the state of affairs existing in the radio-frequency discharge. Electrons do not completely traverse the gap, but oscillate about some mean position, while the positive ions hardly move at all. Thus there is little more interaction between the metal electrodes and the gas molecules with the R.F. discharge on than with it off. A few 721A's have been operated without igniter for as long as 5000 hours with no measurable change in either protection or recovery. This experiment was done at a transmitter power level of 250 kw. peak power. The best life that can be expected with the igniter operating at 100 microamperes is 500 hours, at which time the recovery time is badly deteriorated. In order to maximize the life of the tube, the initial gas filling consists of a minimum amount of hydrogen and as much water vapor as may be introduced without causing excessive leakage power (see Figs. 20 and 23).

Manufacturing and Testing

Some interesting problems occur in the manufacture of copper-disc seal TR tubes which are quite different from those encountered in the construction of more conventional tubes. The copper-disc seals are usually made by high-frequency induction heating. Close control of the spacing between discs must be maintained during the bulb-making operation in which the discs are fused to the glass parts of the tube. One way of accomplishing this is shown in Fig. 34, which depicts a machine setup for making the 724B TR tube. The parts are held by lavite forms which support and locate them during the bulb-making process. The seal is made possible by a correct choice of copper thickness. The copper disc is stressed due to forces set up by the different expansion coefficients of the glass and the copper, and if too thick will pull the seal apart. If too thin, the copper itself will tear. Nevertheless, a properly designed copper-disc seal is very strong; the copper-disc seal TR tubes will pass the JAN1-a* mechanical and thermal shock tests for glass tubes without any difficulty.

The electrical pretuning operation, referred to earlier, comes right in the middle of the manufacturing process. Before the igniter is sealed in, the bulb is placed in a special pretuning cavity. The setup includes an oscillator

* Joint Army-Navy Specification for electron tubes.

of appropriate frequency range, a wavemeter, and some device for indicating resonance. The part inside the glass of one of the copper discs is bent, by gentle tapping, until the resonant frequency of the bulb in the special cavity is within the required tolerance (which may be as small as 0.25%) of the pretuning frequency.

No heat treatment of any kind is used in the pumping of TR tubes. It is obviously unnecessary to subject the tubes to the usual baking, the principal purpose of which is to remove water vapor film from the tube parts. On

Fig. 35—Pump Station for the 724B Vacuum Tube

the contrary, it is difficult to control the water vapor pressure in tubes which have been baked as the parts absorb a surprising quantity of water. The tubes are filled to fairly high pressure, so a diffusion pump is not necessary. Fig. 35 shows a pump station used in production of the 724B.

The test procedure for TR tubes must verify that each individual tube will fulfill its fourfold function of protecting the crystal, of recovering rapidly, and of introducing neither excessive high-level loss nor excessive low-level loss. The tuning of each tube must be verified, and it must pass mechanical and dimensional tests. Fortunately a tube which is otherwise sound will never introduce excessive high-level loss, so no specific test is required.

The protection test may be made either with a c-w oscillator of suffi-

ciently high level to ionize the TR tube gap, or with a magnetron in the equivalent of a radar microwave head. The high-level test bench used for the 724B is shown in Fig. 36. This bench uses a radar microwave head fitted to special plumbing. In either case, the leakage power is measured at a specified R.F. level. For production testing, an actual measurement of recovery time is not used. Instead a test which measures the quantity of

Fig. 36—High Level Test Bench for the 724B Vacuum Tube

water vapor in the tube in a relative way has been developed. This test involves touching some part of the tube envelope with a piece of carbon dioxide ice, so that most of the water vapor is frozen out forming a small spot of ice on the inside of the tube. Only the hydrogen remains, and the resulting change in either the leakage power or the igniter arc drop is indicative of the quantity of hydrogen and of water vapor in the tube. Careful correlation must be made between this simple dry-ice test and absolute recovery time measurements; experience has shown the dry-ice test to be reliable and in the hands of a skilled operator very informative.

Low-level loss and tuning are checked at such low level that the gas does

not ionize in a cavity of restricted tuning range. Every tube must resonate within the range of the tuning adjustment, and the transmission loss through the test cavity must not be excessive.

Two additional tests are made at the time the d-c igniter characteristics are checked. One, igniter interaction, is important in the 721A, the 10 cm. TR tube. This tube has rather large openings in its cone tips, so that if the igniter electrode is sealed in too close to the cone tip, the glow discharge which surrounds it may extend out into the gap. Such a defective tube will show igniter interaction; the low-level loss through it will be more when the igniter arc is on than when it is off. Normal tubes do not show this effect.

The 724B 3 cm. TR tube has such a tiny opening in its cone tip that igniter interaction does not occur. The tube is more subject to igniter oscillations, perhaps because it is filled to a higher gas pressure. The consequences of these oscillations was explained in the section on The Spike. Igniter oscillations are usually due to an improper gas filling, and are detected by means of a cathode ray oscilloscope.

The above tests are made on each tube as it comes off the production line. Some additional tests are made on selected samples to insure that the quality is being maintained. Selected tubes are subjected to mechanical shock tests and to temperature variation tests to verify both their resistance to thermal changes and that their temperature coefficient of frequency is not excessive. Absolute recovery time, Q, and leakage power tests are made on these tubes, and some are set aside for life testing. By all these tests the important electrical properties of the tube are under constant scrutiny and the danger of shipping defective tubes is minimized. The importance of adequate testing can hardly be over-emphasized, as a defective TR tube may render a whole radar system inoperative.

Bell Telephone Company, while at the Bell Laboratories, was responsible for the construction of test equipment and for most of the recovery time measurements reported in this paper. At various periods during the development work Messrs. A. B. Crawford, V. C. Rideout, G. M. Eberhardt and J. P. Schafer were closely associated with the measurement of the system performance of TR tubes. Mr. R. M. Purinton of the Bureau of Ships deserves much credit for his encouragement and assistance in the standardization program which led to the adoption of the 721A and 724B tube designs by all manufacturers. The Thermionics Branch of the Evans Signal Laboratory provided the bulk of the electrical standardization, calibration and engineering service associated with these tubes and assisted in the development of improved test methods. The magnificent production job done by the Western Electric Company and by other manufacturers, particularly by the Sylvania Electric Products Inc. in making these tubes available to the armed services also deserves mention. Perhaps the final mention should go to the many circuit design engineers both within the Bell Laboratories and elsewhere who handled the many difficult problems relating to the design and use of TR cavities in actual radar systems.

APPENDIX A

Analysis of the Idealized TR Box

Schelkunoff has shown* that the impedance of a resonant cavity can be represented in terms of its resonant frequencies as

$$Z = \sum_a \frac{Z_a}{j\left(\dfrac{\omega}{\omega_a} - \dfrac{\omega_a}{\omega}\right) + \dfrac{1}{Q_a}} \tag{1}$$

or in the vicinity of any single resonance as

$$Z = Z_1 + \frac{Z_n}{j\left(\dfrac{\omega}{\omega_n} - \dfrac{\omega_n}{\omega}\right) + \dfrac{1}{Q_n}}. \tag{2}$$

Under most conditions the Z_1 term is negligibly small. We are therefore justified in thinking of the generalized resonant cavity used as a TR switch as a shunt resonant circuit to which are coupled input and output circuits. For the moment we will consider (1) that these external circuits are resistive only, (2) that the Z_1 in equation (2) is zero; and (3) we will restrict the analysis to the in tune condition.

When the cavity is excited by energy supplied from the input circuit there exists in the cavity a certain amount of reactive power which will be

* S. A. Schelkunoff, "Representation of Impedance Functions in Terms of Resonant Frequencies," *Proc. I.R.E.*, vol. 32, pp. 83–90, February (1944).

designated by the symbol P_0. Of this power, a certain fraction δ_0 is dissipated as losses in the cavity itself where

$$\delta_0 = \frac{1}{Q_0}. \tag{3}$$

The symbol Q_0 with the subscript is further defined as the intrinsic Q, that is, the Q without external loading, to differentiate it from the more general Q_L which is the measured Q when the cavity is loaded down by external coupling. It should be noted that this definition of δ differs from the logarithmic decrement by a factor π.

When coupled to the external circuits the loaded δ is increased. On the assumption that the loading effects of the input and output irises are independent we can write

$$\delta_L = \delta_0 + \delta_1 + \delta_2 \tag{4}$$

where δ_L is the loaded δ, and δ_1 and δ_2 are respectively the input and output loadings. Physically the assumption underlying this expression is that the distribution of electromagnetic fields within the cavity is not seriously altered by the input and output coupling devices. This assumption should certainly be valid as long as the absolute values of the δ's are very small compared to unity. Since the δ's usually encountered are of the order of 10^{-3} or less, the assumption seems to be justified.

Equation (4) may be written

$$\frac{1}{Q_L} = \frac{1}{Q_0} + \delta_1 + \delta_2 \tag{5}$$

The values of δ_1 and δ_2 evidently depend upon the ratio of the apparent series resistance which the external coupling introduces into the resonant cavity to the effective reactance of the cavity, that is,

$$\delta_1 = \frac{k_1^2 R_1}{X} \tag{6}$$

where k_1 is the transformation ratio of the input coupling device, R_1 is the resistance of the input circuit and X is the cavity reactance. Similarly

$$\delta_2 = \frac{k_2^2 R_2}{X}. \tag{7}$$

The values of the δ's may be equally well considered as the ratios of the coupled conductance to the shunt susceptance of the cavity considered as a shunt resonant circuit so that equations (6) and (7) become

$$\delta_1 = \frac{G_1}{k_1^2 B} \tag{8}$$

and

$$\delta_2 = \frac{G_2}{k_2^2 B} \qquad (9)$$

when the R's and X are replaced by their reciprocals and transformed from a shunt to a series circuit.

The equivalent circuit is shown in Fig. 37, where for convenience everything is referred to the cavity and the sources for receiving and transmitting are represented by constant current generators, I and I_m respectively.

The Low-Level Transmission. We are now in a position to express the low-level transmission of the cavity. For this purpose we will assume that
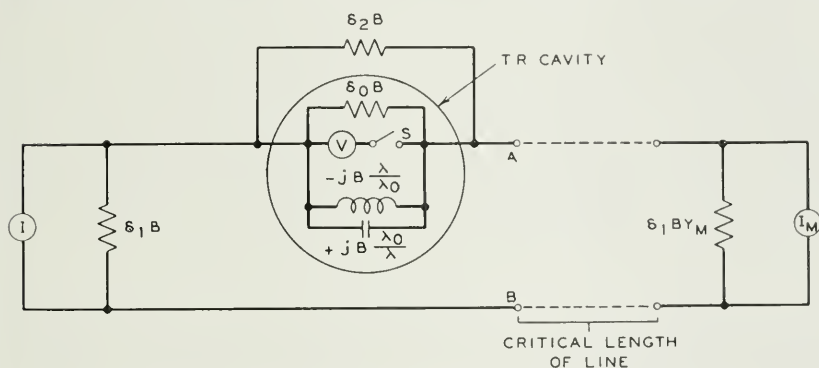
Fig. 37—Equivalent circuit of a system referred to the TR cavity

the admittance of the transmitting branch at plane AB is infinite. The available power is given by

$$P_{\text{avail}} = \frac{I^2}{4\delta_1 B}. \qquad (10)$$

while the power actually going into the load is given by

$$P_{\text{out}} = \frac{I^2 \delta_2 B}{(\delta_0 + \delta_1 + \delta_2)^2 B^2} \qquad (11)$$

the power transmission ratio defined as T is given by

$$T = \frac{4\delta_1 \delta_2}{(\delta_0 + \delta_1 + \delta_2)^2}. \qquad (12)$$

One additional expression is desired. This is the ratio of cavity input resistance to the resistance of the input circuit. This is evidently the reciprocal of the conductance ratio and is given by

$$\sigma = \frac{\delta_1}{\delta_0 + \delta_2}. \qquad (13)$$

The symbol σ is used to call attention to the fact that the in tune impedance ratio is numerically equal to the voltage standing wave ratio on the input line.

The low-level behavior of the cavity is thus defined by three equations.

$$\delta_L = \delta_0 + \delta_1 + \delta_2 \tag{4}$$

$$T = \frac{4\delta_1 \delta_2}{(\delta_0 + \delta_1 + \delta_2)^2} \tag{12}$$

$$\sigma = \frac{\delta_1}{\delta_0 + \delta_2}. \tag{13}$$

High-Level Operation. The high-level performance of the cavity containing a gas discharge can be expressed directly in terms of our original definitions. Fig. 37 still applies, the transmitter admittance changing to its operating value which is assumed to be $\delta_1 B$ at the plane AB. When the gas discharge becomes conducting, the switch S is closed, the value of the reactive power in the cavity (P_0) is set by the character of the discharge and the leakage power is given by

$$P_r = P_0 \delta_2. \tag{14}$$

A constant value of P_0 is equivalent to a constant value of V in the figure. The power dissipated in the cavity walls, the gas discharge and in the output circuit must evidently be given by

$$P_1 = \frac{I_m V}{2} = (P P_0 \delta_1)^{1/2} \tag{15}$$

if $$V << I_m/\delta_1 B.$$

Of this power an amount called the excitation power

$$P_e = P_0 \delta_0 \tag{16}$$

is lost in the cavity walls. The net loss of power in the gas discharge is given by

$$P_g = P_1 - P_r - P_e \tag{17}$$

or

$$P_g = (P P_0 \delta_1)^{1/2} - P_0(\delta_0 + \delta_2). \tag{18}$$

Since the last term is usually very small compared to the first term, we may write

$$P_g \doteqdot (P P_0 \delta_1)^{1/2}. \tag{19}$$

This equation was used for plotting Fig. 16, where δ_1 is replaced by its equivalent in terms of σ, Q_0 and T.

The Derived g Parameters. For some purposes it is convenient to eliminate δ_0 from the expressions for T and σ. This may be done by defining

$$g_1 = \frac{\delta_1}{\delta_0} \tag{20}$$

and

$$g_2 = \frac{\delta_2}{\delta_0}. \tag{21}$$

Introducing these new parameters the equations become

$$\frac{Q_0}{Q_L} = 1 + g_1 + g_2 \tag{22}$$

$$T = \frac{4g_1 g_2}{(1 + g_1 + g_2)^2} \tag{23}$$

$$\sigma = \frac{g_1}{1 + g_2} \tag{24}$$

$$P_r = P_e g_2 \tag{25}$$

$$P_g = (P P_e g_1)^{1/2} \tag{26}$$

The g parameters are particularly useful in defining the behavior of a tube and cavity combination when δ_0 is a fixed quantity while the effects of changes of δ_0 are more clearly shown when the δ parameters are used. The g parameters may be determined experimentally, using equations (21) and (22) without knowing the value of δ_0, that is of Q_0. On the other hand the g's are altered if a tube is replaced by one giving a different Q value while the δ's are intrinsic properties of the coupling mechanisms and remain fixed as long as the cavity and the tube tune at the same frequency and have the same effective reactance.

Tabulation of Related Equations. In the interest of completeness a number of the more important combinations of the basic equations are listed in Table 1. Some of these are of interest for measurement purposes while others apply particularly to actual system conditions.

Off-Resonance Analysis. The analysis can be extended to predict the transmission when the cavity is detuned from resonance by introducing the necessary susceptance term in equation (11) above and solving for T. This gives for the absolute value (neglecting phase)

$$T = \frac{4\delta_1 \delta_2}{[\delta_0 + \delta_1 + \delta_2]^2 + \left[\dfrac{\omega}{\omega_0} - \dfrac{\omega_0}{\omega}\right]^2} \tag{27}$$

TABLE 1.—Relations between Cavity Parameters

		General Expressions		Special Cases	
Quantity	Symbol	In terms of δ's & Q_0	In terms of σ, T, & Q_0	TR case σ = 1	ATR case T = 0
Low level parameters					
Input standing wave ratio	σ	$\dfrac{\delta_1}{\delta_0 + \delta_3}$	σ	1	$\dfrac{Q_0}{Q_L} - 1$
Low Level Transmission	T	$\dfrac{4\delta_1\delta_2}{(\delta_0 + \delta_1 + \delta_2)^2}$	$\dfrac{4\sigma}{(1+\sigma)^2}\left[1 - (1+\sigma)\dfrac{Q_L}{Q_0}\right]$	$1 - 2\dfrac{Q_L}{Q_0}$	0
Q ratio	$\dfrac{Q_0}{Q_L}$	$(\delta_0 + \delta_1 + \delta_2)Q_0$	$\dfrac{4\sigma(1+\sigma)}{4\sigma - (1+\sigma)^2 T}$	$\dfrac{2}{1-T}$	$1+\sigma$
Input δ	δ_1	δ_1	$\dfrac{4\sigma^2}{[4\sigma - (1+\sigma)^2 T]Q_0}$	$\dfrac{1}{(1-T)Q_0}$	$\dfrac{\sigma}{Q_0}$
Output δ	δ_2	δ_2	$\dfrac{(1+\sigma)^2 T}{[4\sigma - (1+\sigma)^2 T]Q_0}$	$\dfrac{T}{(1-T)Q_0}$	0
High level parameters					
Leakage power	P_r	$P_0\delta_3$	$\dfrac{P_0(1+\sigma)^2 T}{Q_0[4\sigma - (1+\sigma)^2 T]}$	$\dfrac{P_0 T}{Q_0(1-T)}$	0
Gas discharge power	P_g	$[PP_0\delta_1]^{\frac{1}{2}} - P_0(\delta_0 + \delta_2)$	$\dfrac{2\sigma}{1+\sigma}\left[\dfrac{PP_r}{T}\right]^{\frac{1}{2}} - \dfrac{4P_r\sigma}{T(1+\sigma)^2}$	$\left[\dfrac{PP_0}{(1-T)Q_0}\right]^{\frac{1}{2}} - \dfrac{P_0}{(1-T)Q_0}$	$\left[\dfrac{PP_0\sigma}{Q_0}\right]^{\frac{1}{2}} - \dfrac{P_0}{Q_0}$

where ω_0 and ω are respectively the resonant angular frequency and the operating angular frequency.

This may be rewritten as

$$T = \frac{T_0}{1 + Q_L^2 \left[\dfrac{\omega}{\omega_0} - \dfrac{\omega_0}{\omega}\right]^2} \tag{28}$$

where T_0 is the in tune transmission and Q_L is the loaded Q, if one assumes that the δ's and Q_L remain unchanged for small departures from the resonant wavelength.

The input impedance of the cavity in terms of the input line impedance is then

$$Z = \frac{\delta_1}{j\left[\dfrac{\omega}{\omega_0} - \dfrac{\omega_0}{\omega}\right] + \delta_0 + \delta_2}. \tag{29}$$

The effect of other resonant modes which have been neglected in this analysis may be included by the addition of a term (σ_1) in equation (29) giving

$$Z = \sigma_1 + \frac{\delta_1}{j\left[\dfrac{\omega}{\omega_0} - \dfrac{\omega_0}{\omega}\right] + \delta_0 + \delta_2}. \tag{30}$$

Equation (30) and equation (2) are identical except for terminology.

APPENDIX B

EXPERIMENTAL DETERMINATION OF "g" PARAMETERS OF WINDOWS

The derived g-parameters which express the electrical size of a window between a resonant cavity and a surge impedance line were defined in Equations 20 and 21 of Appendix A. Numerical values of these parameters may be of some interest, together with their relation to physical dimensions of the windows. The 721A test cavity was used for an experimental determination of the relation between window width and "g." This cavity is $2\frac{1}{16}$ inches inside diameter, and is coupled by means of windows to two $\frac{9}{16}$ diameter coaxial lines. The width of the windows may be adjusted by rotating the coaxial lines so as partially to close the openings. The insertion loss through the cavity was measured at 3100 mc. by means of a superheterodyne receiver which included a calibrated attenuator in its intermediate frequency section. The windows were carefully maintained geometrically equal. In this case,

$$g = \frac{T^{1/2}}{2(1 - T^{1/2})}$$

which follows immediately from equation 23 of Appendix A on the assumption that $g_1 = g_2$. Fig. 38 shows the results; g proves to be proportional to the fifth power of the window width, over a very large range of values of g. A knowledge of this relationship permits one, with the aid of equations 23, 24, 25, 26, and 27 of Appendix A to calculate the window size
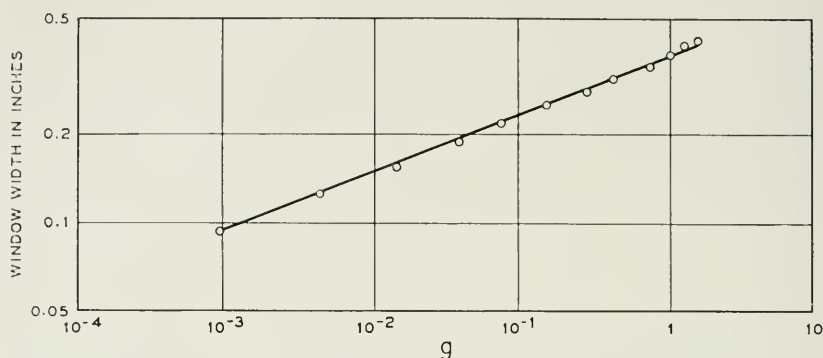
Fig. 38—The relationship between window conductance (g) and window width for the 721A test cavity

necessary to give any desired conditions of match, insertion loss, and leakage power.

APPENDIX C

The ATR Box

The value of the input impedance of the ATR is given by equation 29 with δ_2 equal to zero so that

$$Z = \frac{\delta_1}{j \left[\dfrac{\omega}{\omega_0} - \dfrac{\omega_0}{\omega} \right] + \delta_0} \tag{31}$$

which reduces to

$$Z = \frac{\delta_1}{\delta_0} \tag{32}$$

for the in tune case. This impedance is in series with the magnetron branch and hence restricts the possible range in values for the impedance at plane AB. Defining as F the fraction of the available power which is not absorbed by the ATR, then from Fig. 39 with the admittance of the receiver branch assumed to be $\delta_1 B$,

$$F = \frac{4}{(2 + G)^2 + B^2}, \tag{33}$$
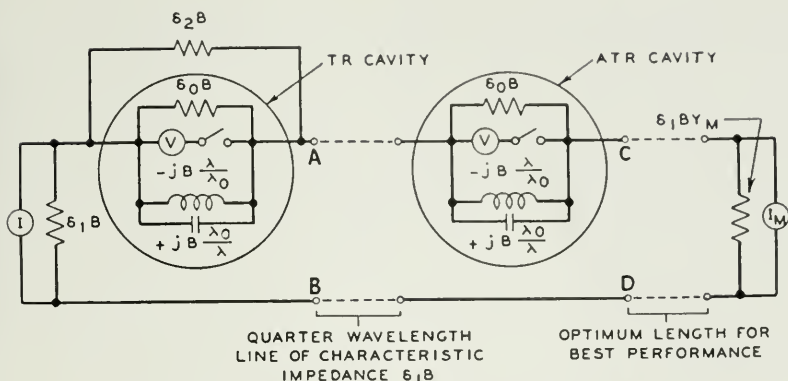
Fig. 39—Equivalent circuit of a system including an ATR
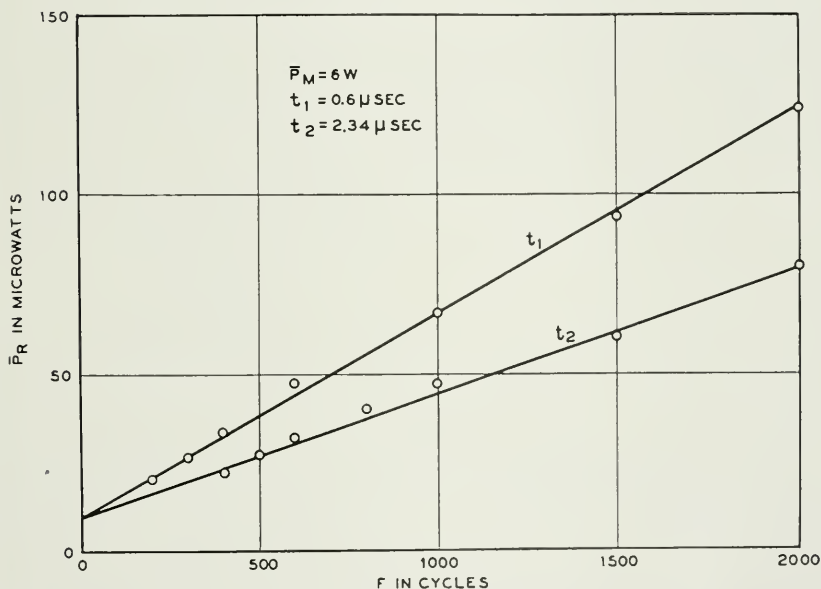
Fig. 40—Average leakage power as a function of repetition rate for two different values of pulse duration for 724B tube

where

$$G - jB = \frac{1}{(Z_m + Z)} \qquad (34)$$

and Z_m is the impedance of the transmitter referred to the cavity and measured at the plane CD. The worst condition will occur when $Z_m = 0$. Under these conditions but assuming that the ATR is in tune

$$F = \frac{4}{(2 + G)^2}, \qquad (35)$$

But now G is the reciprocal of the Z of equation (32) so that

$$F = \frac{4\delta_1^2}{[2\delta_1 + \delta_0]^2}.$$ (36)
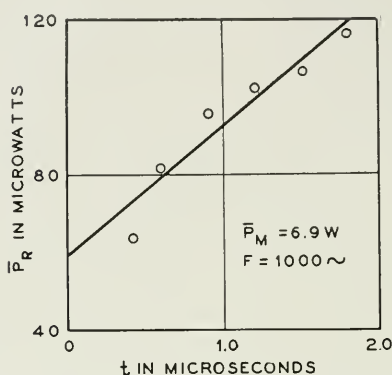
Fig. 41—Average leakage power as a function of pulse duration for the 724B tube
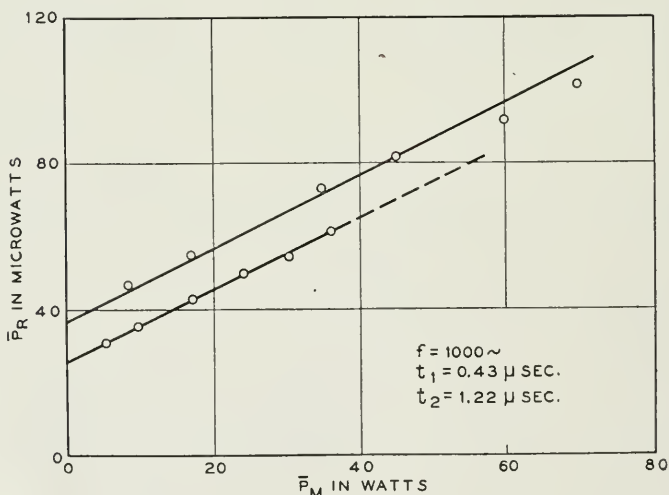
Fig. 42—Average leakage power as a function of average magnetron power for two different values of pulse duration for the 724B tube

This equation is analogous to equation (12) of Appendix A. At high levels the gas discharge power will be given by equation (19) as for the TR

$$P_g = [P\, P_0\, \delta_1]^{1/2}$$ (19)

If the ATR and the TR are designed to have the same value of P_g then the values of δ_1 and δ_0 must be the same so that a relationship will exist between F and T given by

$$F = \frac{4}{(3 - T)^2} \tag{37}$$

The input impedance Z to an ATR adjusted to the same gas discharge power of a TR with a transmission of T is given by

$$Z = \frac{1}{1 - T}. \tag{38}$$

APPENDIX D

The Analysis of Leakage Power Data

The section on receiver protection described the three components of leakage power which were referred to as spike, flat, and direct coupling. One may write down at once the following simple expression for leakage power:

$$\overline{P}_R = E_s f + P_F f t + \overline{P}_M T_D \tag{39}$$

where $\overline{P}_R$ is average leakage power

E_s is energy in a single spike

f is pulse repetition frequency

P_F is flat power

t is pulse duration

$\overline{P}_M$ is average magnetron power (averaged over the recurrence period) and

T_D is direct coupling insertion loss.

Experimental curves verifying the linear relationships indicated by this simple equation are shown in Figs. 40, 41, and 42. It is a straightforward operation to deduce numerical values for the three TR box leakage parameters from the slopes and intercepts of these curves.

Equation (39) was written on the assumption that gas-limited flat power and direct coupling power add linearly. If instead we assume that a phase angle θ exists between the two currents, we find:

$$\overline{P}_R = E_s f + P_F f t + T_D \overline{P}_M + 2\sqrt{P_F f t T_D \overline{P}_M} \cos\theta \tag{40}$$

This of course is identical with equation (39) except for the $\cos\theta$ term. If $\cos\theta$ is not zero, we no longer expect a linear variation of $\overline{P}_R$ with f, t, or $\overline{P}_M$; the experimental curves demonstrate quite clearly that $\cos\theta$ must vanish, hence θ must equal $90°$.

A Wood Soil Contact Culture Technique for Laboratory Study of Wood-Destroying Fungi, Wood Decay and Wood Preservation

By JOHN LEUTRITZ, JR.

Limitations imposed by other biological test methods have largely been overcome by using autoclaved top soil for the substrate and pure cultures of the decay organisms. The use of soil was the direct result of observations on the rapid decay of wood in contact with soil in laboratory termite colonies.

Development of a wood-soil contact culture technique as a result of these observations furnished an excellent laboratory tool for further research on the biological factors promoting and the preservative compounds proposed for preventing decay. Research on the factors promoting decay showed not only that the average top soil furnishes nutrients and nutrilites in the quantity and proportion highly favorable to many decay organisms but also an effective means of regulating the water content of wood or cellulose during the decay period.

Comparisons between laboratory and field results showed the amount of decay obtained by the wood soil contact technique to be more rapid and uniform than decay in the field. The severity of the exposure in the laboratory ensures immediate eliminations of compounds unworthy of further more expensive field studies and evaluates compounds in the same order of effectiveness.

Comparisons and evaluation of wood and cellulose preservatives plus artificial weathering cycles followed by exposure to the method will provide valuable information on initial toxicity and permanence thereby affording a sound basis for the engineering selection of preservatives for a variety of purposes.

L ABORATORY tests for evaluating fungicides are often used as a means of predicting field results and for investigating the action of cellulose and wood-destroying fungi. Of the several laboratory procedures hitherto devised for these purposes, however, none has been entirely adequate. This has led to incorrect interpretation of laboratory assays of fungicidal compounds, with attendant misapplication of preservatives. The confusion and misunderstanding concerning the use of preservatives have been further increased by the misapplication of the laboratory procedures themselves. A brief review and explanation of some procedures and their application will clarify these statements.

Minute quantities of toxic agents and growth-promoting substances which are not readily detected by known chemical analyses may be determined by bio assay methods, the value of which depends upon a prior determination of the reaction of one or more organisms to known quantities of these substances. Another bio assay is the so-called "acceptance test" for fungicides, by which the fungus resistant qualities of materials impregnated with fungicides may be determined. Since fungus resistant qualities are the primary concern in such a test, the identity and quantity of the preservative are of only incidental interest. However, the identity, fungus-proof qualities and quantity of fungicidal compounds are important when

laboratory procedures are devised for comparing effectiveness in the development of different preservatives. In addition, the chemical and physical properties of the different preservatives must be considered for the determination of their subsequent behavior when exposed to a variety of environmental conditions. Bio assays may thus be used for quantitative, qualitative, comparative, or predictive purposes.

In order to survey existing tests, it may be helpful to classify them. There are three groups of rather ill-defined laboratory methods based on the nutrient and physical properties of the substrate. The first group is comprised of those methods in which an agar or similar base is used. Various nutrients or nutrilites* may be added to this base[1], and prior to inoculation with one or more fungi the preservative may also be added. This group includes the standard petri dish test described by Richards[2], 1923, which has had extensive use in the field of wood preservation. The carbohydrate source in the standard petri dish method was malt sugar. Later, in response to the requests by industry, Richards attempted to substitute wood flour as the nutrient. However, the radial fungus growth used as the criterion of toxicity was very sparse and thin and the substitution of wood flour for sugar was discarded. It is of interest to record here that Richards also summarized the previous work on toximetric tests of wood preservatives.

The second group includes those methods in which the preservative is added directly to a cellulose material before exposure to organisms. The preserved material may be the only source of nutrient for the fungi, or a piece of similar untreated material may be provided. Such a method is described in a paper by Waterman, Leutritz and Hill[3], 1938. No agar is used, and the untreated wood is supported over water by mechanical means. When agar is used to support the preserved material and to supply water, nutrients, nutrilites or combinations of each of these may be added to the agar. This may be done in several ways, among which are the kolle flask method for wood preservatives described by Falck[4], 1927, the standard method of the American Society for Testing Materials for testing fabrics[5], 1942, and the present Signal Corps test of fungicidal coatings[6], 1943. Of these, the first two methods are used chiefly as "acceptance" tests by determining the fungus-proof qualities of fungicidally treated wood and fabrics. They are also used in development work for comparison and for predicting the field behavior of preservatives when supplemented by artificial weathering cycles. The Signal Corps test is used as an acceptance test of fungicidal coatings which are sprayed on electical equipment. Since the criterion is the inhibition of fungus growth at some distance from a paper impreg-

---

* Nutrients here include the sugars and compounds used by the fungi for food purposes, and nutrilites will be referred to in this paper as those compounds necessary to fungus nutrition, such as vitamins, growth substances and minerals, Williams, R. J., 1928. (See Bibliography at end of this paper.)

nated with the fungicidal coating it is fundamentally a quantitative measure of the amount of fungicide which diffuses into the agar from the impregnated paper specimen.

A third group of test procedures employs soil or soil suspension in conjunction with the preservative materials. Here the soil furnishes an active microbial culture and supplementary nutrients and nutrilites. The soil suspension method has been described by Furry, and Zametkin[7], 1943, and the soil burial method by the American Society for Testing Materials.

The techniques included in the first group are time saving, permit of replication, and are readily duplicated by other investigators. However, the results in the agar-fungicide system do not apply to a cellulose-fungicide system and are therefore a source of confusion resulting from their misinterpretation when so used. Agar-fungicide systems as originally described by Richards are quantitative tests and have been used principally for comparative toxicity studies. From such comparative studies attempts to predict the behavior of a preservative in subsequent field tests have been generally unsuccessful. Examples of the discrepancies between the results from field and petri dish tests will be discussed later in this paper.

In general, the second group of methods takes a longer time, and replication leaves much to be desired. Since the preserved material is the same for laboratory and field tests, better agreement between field and laboratory results should be obtained with the kolle flask-wood block method and the A.S.T.M. fabric methods. However, the Signal Corps method for testing fungicidal coatings used on electrical equipment is not a a true test of the coating material per se.

The third group of methods introduces a large number of variables through the use of soil. Previously, replication of results and concomitant duplication by other investigators had been lacking, due to microbial activity, physical properties, nutrient properties, and moisture variations of the soil. However, during experimental work with termites, the author[8] made certain observations on the various factors involved in the decay process. These led to an intensive study of the problem resulting in the development of a test method for wood preservatives which overcomes many of the limitations of earlier methods. The soil burial method is a severe test of fungicide treated material, and, with the modification to be discussed in this paper, it is anticipated that the variables which cause non-uniformity of results can be eliminated. The method is also evaluated by comparison between the results obtained in the laboratory and those obtained from parallel field tests.

Rapid decay of wood in contact with soil was observed during an attempt to establish experimental termite colonies in the laboratory (Leutritz[8], 1939). Instead of becoming infested by the termites, nearly all the blocks

decayed more rapidly and more completely than in any previous laboratory test. Preliminary experiments were devised to ascertain the factors responsible for the accelerated decay and to establish optimum conditions for growth of fungi in laboratory tests of wood preservatives. As a result of this exploratory work a laboratory technique was devised which permitted study of these factors and which offered a convenient means of evaluating toxicity and preservative properties of chemical compounds. Further investigation was made on the effect of nutrients and nutrilites in the soil, temperature, and the moisture content of the wood. Parallel with this laboratory investigation, a study was made of the fungus attack on wood under climatic conditions very favorable for decay at Gulfport, Mississippi.

Initial Experiments and Results

As a preliminary step, the moisture content of the soil from the termite colonies was determined by oven-drying 100-gram samples. This was found to average 22% of the oven-dry weight of the soil. Tests with several soils showed that approximately the same moisture content could be obtained by merely adding to dry soil just enough water to make the mixture cohere when squeezed in the hand.

A one-hundred-gram sample of moist soil was placed in each of 24 large-mouthed, eight-ounce, screw-capped bottles (12 cm. high and 6 cm. in diameter). A weighed oven-dry block of southern pine sapwood, 2 x 2 x 2 cm., was pushed to a depth of about 2 cm. into the soil in each bottle. The caps were put on, and the preparations were sterilized for 30 minutes at 15 pounds' pressure in an autoclave. After cooling, the block in each of twelve of the bottles was inoculated with a pure culture of one of seven common wood-destorying fungi—*Lentinus lepideus, Fomes roseus, Poria microspora,** *Polyporus vaporarius, Coniophora cerebella, Poria incrassata,* and *Lenzites trabea.* Twelve bottles, not inoculated, were used for moisture determinations.

The bottles were then placed in an incubator maintained at 26°–28°C. At the end of each month three of the bottles inoculated with each fungus were taken from the incubator. Each block was removed from the soil and weighed immediately; it was then allowed to dry in an oven at 105°–110°C. to a constant weight. The average percentage loss in dry weight due to decay was calculated. The results, recorded in Fig. 1, show that the very rapid decay of wooden blocks in contact with the soil is not the result of any one particularly active fungus. Each of the seven species produced exceedingly rapid decay under the conditions of the soil assay.

* This fungus was designated BTL U-10 until recently identified as *Poria microspora* by Miss Mildred K. Nobles, Dept. of Agriculture, Ottawa, Canada, 1943. (See Bibliography at end of paper.)

Fig. 1—Wood soil contact technique

For comparison, a similar test was made according to the method described by Waterman, Leutritz, and Hill[3], 1938, in which the test blocks are placed on inoculated sapwood slabs supported over water in capped wide-mouthed bottles. Comparison of the average percent weight loss due to decay for all organisms by both methods, Fig. 2, shows that the water-wood method is far less effective in producing decay than the soil method.
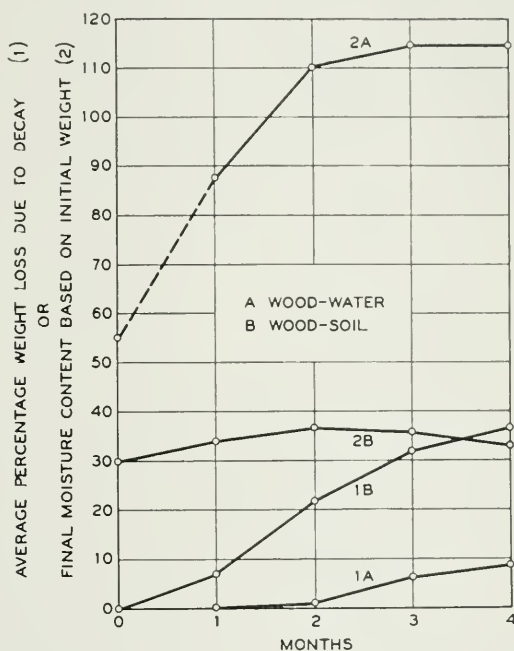
Fig. 2—Comparison of average weight loss and final moisture content by wood-water and wood-soil techniques

An additional experiment was conducted with several strains of two of the fungi previously used, *Coniophora cerebella* and *Lentinus lepideus*. The *Coniophora cerebella* strains were as follows:

Baarn, from Dr. Johanna Westerdyjk, Holland

Liese, from Dr. Liese, Germany

Idaweiche, from Dr. Idaweiche, Germany

Madison, from Forest Products Laboratory, Madison, Wisconsin, isolated from oak, November 13, 1919

BTL, also from Forest Products Laboratory, Madison, Wisconsin, 1930

The *Lentinus lepideus* strains were from the following sources:

No. 534, from Forest Products Laboratory, Madison (No. 534)

BTL U-1, U-13, U-14, and U-32, from creosoted pine telephone poles which had failed in service

Gulfport, from a test post in Gulfport, Mississippi, used for our assay work

The results of the assay with these strains of *Coniophora cerebella* and *Lentinus lepideus* showed the average weight loss in percent due to decay to be 32.0 and 27.3 percent respectively which was as great as that in the previous soil tests with the single representative of these species. A greater amount of decay was obtained with one strain of Coniophora cerebella due to a slight change in technique, i.e., the fungus was first established on small slabs of southern pine sapwood, and then sterile oven-dry blocks were dropped on the vigorously growing fungus. The large amount of decay (60%) which resulted led to the adoption of this modification in all subsequent tests.

The foregoing tests may be regarded as supporting the use of the criteria previously employed in the selection of fungi for laboratory tests—namely, their occurrence as saprophytes of wood, their isolation from service materials for example, pine telephone poles or tests posts, and their demonstrated ability to bring about decay of wood in the laboratory.

As a result of these preliminary experiments, the use of soil as the medium in testing procedure was adopted.

Soil Contact Technique

On the basis of the foregoing experiments and in view of the rapidity of the decay occurring on test blocks in the soil-contact test, the following method is described as a means of evaluating the effectiveness of preservatives or toxic materials which are recommended for the protection of wood or other cellulosic materials. The method may also be used to study environmental factors which affect decay or it may be adapted to the study of fungi other than the wood-destroying fungi of the Basidiomycetes.

Ordinary top soil, such as a florist would use for potted plants, is satisfactory for the test. While experience has shown that top soil from a number of different sources may be used without materially affecting the results, standardization would be desirable. Therefore the term "soil" will be defined as a sandy loam type which contains 4–6 percent of organic matter and a pH originally between 5–7. The soil is passed through a coarse-mesh screen to remove rubble, stones and other debris; this is most easily accomplished when the soil is dry. The screened soil is moistened with just enough water to effect cohesion into a soft ball when squeezed in the hand, and a check may then be made by determining the moisture content of the soil. When prepared in this manner the moisture content of the soil should be 20–25% on an oven-dry weight basis. In an alternative procedure, the moisture content of the dry soil is ascertained and then sufficient water is added to give a moisture content of 20–25%.

Bottles, 12 cm. high and 6 cm. in diameter, are half filled (60–100 grams)

with the moistened soil. Two pieces of southern pine sapwood "feeder strips" (3.5 x 2.0 x 0.3 cm.) are placed on the soil in each bottle, Fig. 3. The bottles are closed tightly with screw caps and then autoclaved for 30 minutes at 15 pounds' pressure. When the bottles have cooled, a small inoculum (a few millimeters square) cut from a pure culture of a suitable wood-destroying fungus is placed on the sapwood substrate. Each bottle contains a single dominant fungus culture. It is best to use at least four to eight selected species of fungi for an assay. The bottles are again capped and placed in an incubator, or a controlled temperature room, held at 26°– 28°C., for at least one month. Any contaminated or weak cultures are
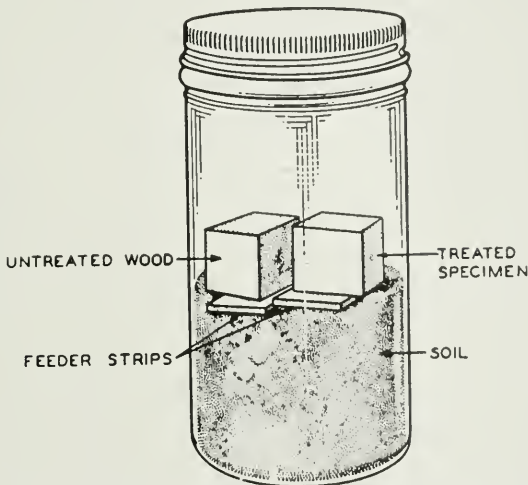
Fig. 3—Schematic diagram of wood soil contact method

discarded. This completes the preparation of the pure fungus cultures, Fig. 4, and they are now ready to receive the test blocks.

To each bottle containing a culture established on sapwood substrate are added an untreated control block and a block treated with a preservative according to the following method:

The required number of $\frac{3}{4}''$ cubes of sapwood blocks are placed in a humidity chamber at 30°C. and 76% relative humidity until the blocks have reached a constant weight. Then the necessary number of weighed blocks, weighted to ensure immersion, are placed in a container of convenient size under a bell jar fitted with a separatory funnel. After evacuation of the bell jar to a pressure not greater than 2 cm. as measured by a mercury manometer, the vacuum is held for 5 minutes. The stopcock in the pump line is then closed, and sufficient solution is admitted from the separatory funnel

Fig. 4—Pure fungus culture of Poria incrassata

to submerge the blocks completely when the air is admitted. After remaining in the solution for 5 minutes, the blocks are wiped superficially and and weighed. This treated weight is used for calculation of the theoretical retention according to the following formula:

$$R = \frac{GC\,(62.4)}{100\ V}$$

in which R^* = pounds of preservative per cubic foot of wood, G = gain in weight in grams, C = grams of the preservative in 100 grams of solution, and V = volume of the test piece in cubic centimeters. When the solvent has evaporated from the blocks, they are placed on racks, returned to the humidity chamber and brought to constant weight. The difference between the humidity weights before and after treatment serves as the basis for calculating the actual retention, and the final equilibrium weight is used also as the initial weight of the treated block before exposure to the fungus.

The cross-section side of the blocks is placed in contact with the vigorously growing mycelia of the sapwood culture which in turn is in contact with the soil. For each concentration of preservative, three treated blocks and three untreated control blocks are exposed to each species of fungus. The bottles are recapped and place in an incubator or constant temperature room at $26°-28°$ C. with a relative humidity of $85-95\%$.

Exposure of the blocks to the fungus for from twelve to twenty-four weeks gives satisfactory results. If a sufficient number of treated specimens is exposed to the same organism, one or two specimens may be removed at the end of twelve weeks; and if considerable decay has taken place, the test may be concluded. At the end of the exposure period the blocks are brushed free of mycelia and immediately weighed to determine their moisture content. The blocks are then allowed to stand in the room until dry, after which they are again transferred to the humid chamber (temperature $30°C.$, relative humidity 76%) for two-three days until a constant weight is attained.

In the toxicity work reported in this paper an untreated reference block was added to each test bottle. If a large number of assays are contemplated, the number of weighings may be reduced by eliminating the untreated blocks except for occasional reference purposes. The reduction in the number of reference blocks may be accomplished by establishing a decay norm for each test organism. This norm would be based on data similar to that used in Fig. 1 except that the procedure would be the same as that described for treated blocks. Comparison of the percentage weight loss due to decay

* To express the retention metrically $R \times 16.018 = \dfrac{\text{Kilograms}}{\text{Cubic meter}}$.

of the norm with the percentage weight loss due to decay of the test block may be used as a measure of the effectiveness of the preservative or toxic material. An index of the value of a preservative treatment may be obtained from the following computation:

$$\frac{\%\ \text{loss of norm} - \%\ \text{loss of treated block}}{\%\ \text{loss of norm}} \times 100$$

Values of the index would range from 100, representing complete protection against decay, to 0, representing no protection whatever.

In most cases, especially when no volatile preservative is present, the untreated reference block disintegrates completely within twelve to sixteen weeks. An inspection rating based on strength may be used to supplement weight loss due to decay. The rating is made on the basis of appearance and strength: 10 denotes a sound condition, 9 superficial decay, 8 superficial decay in spots or streaks, 7 general surface decay, 6 considerable decay but not enough to allow specimen to be broken easily, 5 advanced decay, and 4, 3, 2, and 1 different stages of advanced decay, determined primarily by the ease with which the specimen is broken; 0 denotes complete disintegration. Ratings of 5 and below are considered failures. Similar ratings have been used for sticks in field work. This method of rating was used in the field trials of preservatives which appear later in this paper. Although the system is an arbitrary one, considerable correlation has been shown between these dissection ratings and the weight loss in percentage. With a series of field sticks or blocks treated with the same low amount of a preservative, ratings based on strength for the series are found to be very closely correlated with weight losses, even when the ratings are made by different workers.

The Influence of Moisture on Decay

The first experimental factor studied was the moisture content of the wood preceding and during the time that decay took place. Statements in the literature concerning the optimum moisture content for the decay of wood have placed the figure variously from fiber saturation, 27–30% to 60% (Schorger, 1926)[10] and 150% (Benton & Ehrlich, 1940)[11] of the wood substance based on the oven-dry weight.

Figure 1 gives data on the moisture content of blocks exposed to the seven species of fungi for periods of one, two, three, and four months. Uninoculated control blocks, removed at the end of each of these periods, were found to be at fiber saturation, indicating 100% relative humidity in the bottles and little or no migration of liquid water.

During the progress of decay there is a rapid decrease in the weight of the wood substance. But the amount of water present in each block does

not decrease and at all stages of decay corresponds to about 35% of the original dry weight of the block. Since the amount of water does not decrease as the amount of wood substance decreases, there is an increase in the percentage of water expressed in terms of dry weight as shown by curves labeled 3 in Fig. 1. Such an increase in the amount of water relative to the remaining wood substance would tend to limit decay if the optimum moisture content for initiating decay is considered to be at or near fiber saturation.

If the absolute amount of water in the blocks does not change during the progress of decay, then the final amount of moisture divided by the initial weight of the blocks should give a percentage figure fairly close to the initial fiber saturation of approximately 30%. The curves labeled 2 in Fig. 1 showing the moisture content based on the initial weight indicate that this is the case. For example, Fig. 1 shows that the average for all organisms is 4% greater than 30% after one month, 7% greater after two, 6.3 after three, and only 2.1 greater after four months. Between the third and fourth months the soil showed signs of drying out and examination of all of the moisture curves, Fig. 1, indicates a loss of water through the bottle caps, which accounts partially for the discrepancy. The 5–10% increase in water over the original fiber saturation may be due to slight condensation on the blocks or to the respiratory activity of the fungi in breaking down carbohydrates into CO_2 and water.

The average amount of decay (curve 1) and the final moisture content referred to the initial weight (curve 2) for all the organisms obtained by the wood-water (A) and wood-soil (B) assays are compared in Fig. 2. The water content of the blocks in the water test varied from 55–165% based on the oven-dry weight of wood and the decay was much less in amount and uniformity than that obtained by the wood-soil technique with the same organisms. From the results for individual blocks, the limiting water content at which no decay took place was determined as 78% for *Poria incrassata*, 84% for *Coniophora cerebella*, and 66% for *Polyporus vaporarius*. In a few instances, despite the full cell saturated conditions, decay did take place. Examination of these blocks indicated that most of the decay was confined to the surfaces of the blocks. This indicates that lack of oxygen was the limiting factor.

When wood was supported on glass rods over agar (kolle flash technique) full cell saturation of the blocks often occurs due to capillarity of the glass, condensation of water, accidental contact between wood and agar, and conduction by the fungus filaments. That the water content of the wood in the kolle flask technique is also too high is indicated by the "optimum" moisture content of 150% and the relatively small weight losses due to decay, less than 10%, cited in the experiments of Benton and Ehrlich.[11] The amount of decay was again shown to be affected by the water content

of the wood. If decay is to be used as a criterion of toxic effectiveness, the importance of eliminating variations in the water content of the block can be fully realized. The wood soil technique offers an excellent means of controlling moisture for studies of wood decay.

Sand, cotton, sawdust, wood flour, and soils with varying moisture contents were also used as supporting substrates, but in no case was the amount of decay as great as that with the same technique using soil of 20–25% moisture described above. When soils with water contents of 5%, 10%, 20–25% and 30% were compared, the moisture contents of blocks in contact with them were 12.8%, 23.9%, 27–30% and 73.9%, respectively. Decay of the blocks was adversely affected by lack of moisture in the first two cases and by full cell saturation of the wood in the last instance. However, if moisture were the only controlling factor the amount of decay of wood in contact with sand should be comparable to that in soil, but this was not the case.

The Influence of Soil Nutrient or Nutrilites

Nitrogen in the form of asparagine has been shown by Schmitz and Kaufert[12] (1936) to cause an increase in the amount of decay of *Pinus resinosa* by *Lenzites trabea*. Since wood contains only about 0.1% to 0.3% of nitrogen, any additional nitrogen received from the soil should promote decay. It might be expected that the soil supplies nitrogenous and other nutrients, nutrilites, vitamins, etc., that accelerate decay. Evidence for this was obtained by comparing the decay of blocks in contact with (1) top soil, (2) top soil that had been leached for several days with hot water, and (3) three artificial soils composed of washed sand and fuller's earth. In this experiment *Poria incrassata* was used as the inoculum for a test period of 12 weeks. The moisture content of the uninoculated control blocks was 27–30 percent and of the substrate for each series 22 percent; the temperature and time were constant. The average weight loss for the blocks in contact with top soil was 54%, with water extracted top soil 45.4% and the three mixtures of sand-fuller's earth 24.7%.

It is apparent from these data that the top soil promotes decay to a far greater extent than the sand-fuller's earth mixtures and slightly more than the water extracted top soil, despite the same moisture (fiber saturation) content of the blocks. The actual rate of decay of blocks in contact with the top soil was more than double that of the other mixtures. Therefore, the conclusion may be drawn that nutrients or nutrilites are present in the top soil which stimulate growth of the fungus and promote decay.

Since the water-extracted soil proved not so favorable for decay as the original top soil, some of the growth-promoting substances must have been soluble in water. The greater decay of wood in contact with the extracted

top soil than of that on the mixtures of sand and fuller's earth indicates that the nutrients present in the soil were not all removed by the water extraction. Further information was obtained by adding soil extract and other nutrient solutions to the sand-fuller's earth mixtures. The following materials were used:

    4500 grams of washed beach sand
     900 grams of fuller's earth
     300 ml. of soil extract
      60 ml. of malt extract (2% water solution)
      50 leached blocks
         vitamins B_1, B_6 and biotin (free acid)
         stock mineral solution of the following composition:
             1.5 grams per liter of KH_2PO_4
             1.0 gram per liter of $MgSO_47H_2O$
         plus the following elements in parts per million:
             0.02 Cu, 0.01 Mn, 0.005 B, 0.10 Fe, 0.01 Mo, 0.09 Zn
         pure cultures of the fungus *Poria incrassata*
      42 eight-ounce bottles, screw capped (12 cm. high by 6 cm. diameter)
         maltose

Procedure:

The sand and fuller's earth in the proportions mentioned above were mixed in a porcelain jar on a ball mill for six hours. The blocks were leached for two years with weekly changes of distilled water, using 50 ml. of distilled water for each block. Before the test, the blocks were oven-dried to constant weight at 105°C. and a volume measurement was made by mercury displacement method. The volume at fiber saturation was calculated from the oven-dry volume according to the formula: Volume at fiber saturation = Oven-dry volume + 0.25 × oven-dry weight.

One hundred grams of the soil moistened with 20 ml. of the appropriate nutrient solution (Table 1) was placed in an eight-ounce bottle for each block. The weighed block was pushed into the soil with a cross-section of the block facing upward until the top of the block was level with the soil.

The bottles were capped and autoclaved for 20 minutes at 20 pounds' pressure. After sterilization and cooling, an inoculum from a pure culture of the fungus *Poria incrassata* was placed on the top of each block. The bottles were then placed in a controlled temperature room (26°–28°C., relative humidity 90–95%) for 16 weeks.

At the end of this time the blocks, brushed free of soil and mycelia, were weighed immediately, and the volume was measured. Finally, the blocks were again oven-dried to constant weight and the volume was measured again.

The average volume at fiber saturation of the 42 blocks calculated from the initial volume when oven-dry was found to be 7.24 cc., and the final average volume when removed from the test was 7.26 cc. No shrinkage took place until the blocks were oven-dried, and then the distortion became

permanent. This constancy of the volume during the decay period indicates the mechanism by which the water is held practically constant during the decay period. Any loss of water would result in a shrinkage from which there would be no recovery.

Table 1 gives a list of the solutions used to moisten the artificial soil and the average weight loss for 3 blocks in percentage for each variation. Several

TABLE 1

EFFECT ON THE DECAY OF WOOD IN CONTACT WITH SAND AND FULLER'S EARTH
MIXTURES MOISTENED WITH VARIOUS NUTRIENTS AND NUTRILITES
Organism *Poria Incrassata.* Time 12 Weeks

Solution Used to Moisten Artificial Soil	Average Weight Loss in Per Cent Due to Decay
Top Soil (Control)..	67.0
M.S.* + 0.2% Ammonium Nitrate + 1% Maltose + Vitamins B_1, B_6 and Biotin†..	62.7
2% Malt Extract...	59.3
M.S. + 2% Ammonium Nitrate + Vitamins..................	51.3
M.S. + 2% Ammonium Nitrate...............................	47.9
M.S. + Various Combinations of Vitamins‡..................	29.1
Distilled Water..	28.4
M.S. + Vitamins + 1% Maltose.............................	11 6
M.S. + 1% Maltose..	12.9

* M.S. = mineral solution containing the following minerals:
    Potassium Dihydrogen Phosphate.............. 1.5 grams per liter
    Magnesium Sulfate........................... 1.0 gram per liter
    Copper...................................... 0.02 parts per million
    Manganese.................................. 0.01 parts per million
    Boron....................................... 0.005 parts per million
    Iron.. 0.10 parts per million
    Molybdenum................................. 0.10 parts per million
    Zinc.. 0.09 parts per million
† The vitamins used and the concentrations per liter were as follows:
    B_1... 0.1 milligrams per liter
    B_6... 0.1 milligrams per liter
    Biotin....................................... 0.02 milligrams per liter
‡ The vitamins were added singly and in the following combinations:
    B_1 + B_6 + Biotin—Concentration of each vitamin as listed above.
    B_1 + B_6
    B_1 + Biotin
    B_6 + Biotin

conclusions may be drawn. It is evident that the soil greatly accelerates decay, and that the soil extract contains a large portion of the nutrients and nutrilites which accelerate decay. Malt extract, which contains proteins and sugars, and the mineral solution fortified with ammonium nitrate also stimulate decay. The effect of nitrogen in increasing decay confirms the experiments made by Schmitz and Kaufert,[12] 1936. When nitrogen is lacking and a simple sugar is present, the fungus consumes the simpler sugar instead of the more complex carbohydrate cellulose. This preference is not evident if sufficient nitrogen is present since both carbohy-

drates are destroyed. There is a slight indication that the vitamin mixtures promote decay but the effect as measured by the weight loss is not very pronounced. The basic mineral solution which was used in this experiment promoted only slightly more decay than the distilled water. While the results are not included in the table, it may be stated that leaching of the blocks had no apparent effect on decay when compared with unleached blocks.

Influence of Temperature on Decay

Most of the early work in the Bell Telephone Laboratories was conducted by the petri dish method at temperatures in the range 26°–28°C., following the recommendations of Richards, 1923. But certain fungi, including *Merulius lachrymans*, failed to grow at this temperature. When several inocula of *Merulius lachrymans* that had failed to grow at 26°–28°C. were transplanted to sterile blocks, according to the earlier sapwood-water technique, the loss in weight due to decay after six months averaged 34% at 21°C. and only 7% at 26°–28°C.

An experiment was planned to test the influence of a wide range of temperatures on the decay of wood in the soil contact assay method. Four kinds of fungi were established under sterile conditions on untreated wood slabs laid on moist garden soil. An abundant growth of the fungi was secured within one to two months. Cubes of sapwood were placed on the vigorously growing mycelia, both of which had been conditioned by exposure overnight to the various temperatures. Sterile soil, also conditioned to the temperatures, was used to cover the blocks. After 15 weeks' exposure, the results were as follows:

	Average Weight Loss in Per Cent				
	0°C.	21°C.	26–28°C.	30°C.	35°C.
Poria incrassata	0.0	35.6	58.2	34.7	0.7
Polyporus vaporarius	0.0	42.7	58.9	1.0	0.8
Poria microspora	0.0	53.3	59.0	42.4	2.4
BTL-U-11	0.0	27.1	62.7	60.5	1.3

The results indicate that the standard temperature, 26°–28°C., was optimum for the four fungi tested. No decay was produced by any of the fungi at 0°C. A temperature of 35°C. was too high for active decay; in the case of *Poria microspora*, for instance, only a single block was attacked. The series at 35°C. was repeated because the soil in some of the bottles seemed to have become rather dry, although the blocks contained 30% moisture. In the new series the humidity was maintained at 76% around the bottles to reduce loss of water, and three more organisms were used. The results

of the previous temperature tests were confirmed and the weight losses due to decay by the three additional fungi, which are known to tolerate higher temperatures, were as follows:

Organism	Percentage Weight Loss Due to Decay
Lentinus lepideus...	21.8
Lenzites sepiaria..	21.3
Lenzites trabea (BTL U-40).................................	44.0

These results indicate that certain fungi are able to bring about decay of wood over a wider temperature range than others. It is clear that a complete statement cannot be made until the effects of various temperatures between 0°C. and 21°C. have been ascertained. In the light of results with *Lentinus lepideus*, *Lenzites sepiaria*, and *Lenzites trabea* showing considerable decay at 35°C., the upper limits of temperature should be determined for these fungi.

Humphrey and Siggers,[13] 1933, studied the effects of different temperatures on the growth of sixty-four fungi. Two different nutrient substrates were used, but the optimum temperature with these rarely differed by more than 2°C. The following summary shows a comparison of their results with those obtained in the above tests:

	Optimum, °C.		Upper Limit, °C.	
	H&S	BTL	H&S	BTL
Merulius lachrymans..........	20	21*	28	>28
Poria incrassata..............	24–30	26–28	34	34
Lentinus lepideus.............	28	28	36	>35
Lenzites trabea...............	28–36	28–35	40	>35
Lenzites sepiaria.............	28–36	28		>35

* Bottle method used; no test has been made yet with soil.

Two of the fungi, *Merulius lachrymans* and *Lentinus lepideus*, brought about decay at limits higher than those reported for cessation of growth by Humphrey and Siggers. *Poria incrassata* had the same limiting temperature in both tests. The temperatures for maximum growth and maximum decay check rather well in both tests.

Field Studies

The rapid decay obtained in the foregoing laboratory experiments based on the soil technique was further evaluated by investigating the rapidity of decay in the field.

Selection of Wood:

For both laboratory and field assays care is exercised in selecting the wood. Boards of southern pine sapwood of the shortleaf type, which includes *Pinus echinata*, and *Pinus taeda*, are obtained from local lumber dealers and are cut into sticks $\frac{3}{4}$ x $\frac{3}{4}$ x 32 inches. Since the square sticks facilitate calculation of volume and retention of toxics or preservatives, they have superseded the round saplings cited by Waterman and Williams,[14] 1934. The sticks are selected on the basis of uniformity of growth, density and ratio of springwood to summerwood. The presence of any heartwood,
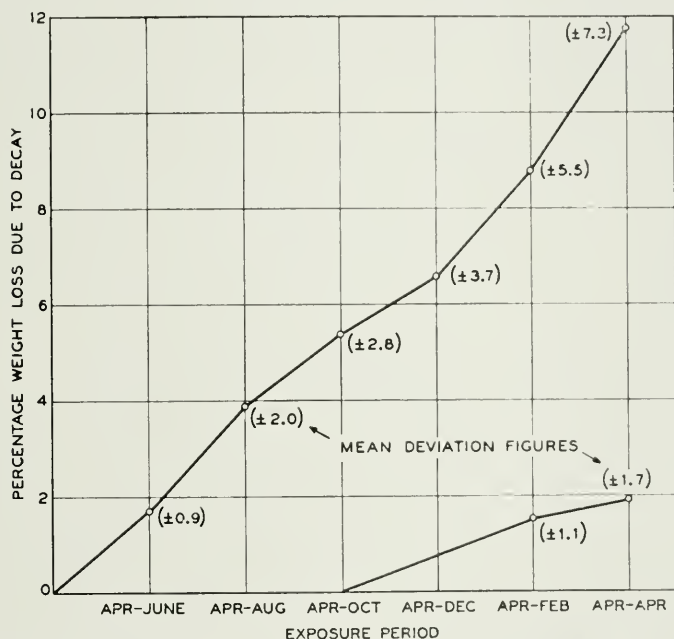
Fig. 5—Field test of untreated sapwood squares exposed at Gulfport, Mississippi, 1941-42

sap stain or other indication of incipient invasion by fungi is cause for rejection. After classification into piles according to arbitrarily chosen weight increments, twenty to twenty-five 32″ sticks, for each concentration of preservative used in field studies are selected by taking the appropriate number of specimens from each pile to give a representative distribution based on density. Since the specimens are subsequently cut in half each individual treatment is represented by 40-50 specimens. Sticks in the median range of density are generally used for laboratory studies after they have been cut into $\frac{3}{4}$″ cubes (8 cc. volume).

An experiment with untreated sticks was carried out at Gulfport, Mississippi, where the climatic conditions are very favorable for decay and also

for termite attack. Six hundred 8-inch lengths were dried in the oven in the laboratory at 105°–110°C. and then weighed. The specimens were then shipped to Gulfport, Mississippi, and distributed throughout the test plot in April. Each eight-inch specimen was buried in the soil until the end of the specimen was even with the level of the soil. At the end of each two-month period subsequent to exposure about 80 specimens were removed, brushed free of dirt and mycelia, oven-dried, and reweighed. In order to study decay during the winter months, 150 additional pieces were planted in October; seventy-five of these were removed after four months and the rest after six months exposure.

From Fig. 5 it is evident that in the field test the loss in weight due to decay was far less than that obtained in the soil test in the laboratory. The maximum amount of decay in the field after two months was as high as 10% in only two out of the 81 samples exposed; after four months it was 19%, after six months 30%, after eight months 20%, after 10 months 30%, and after 12 months 50%. While the maximum percentage loss applies only to one specimen in each case, in general the average percentage losses noted in the figure were far below these figures. Therefore, decay in the field does not approach in uniformity and rapidity that occurring under controlled conditions in the laboratory.

Experiment at Chester, New Jersey, Using Various Nutrients:

Another field experiment was devised in which an attempt was made to increase the rate of decay by using nutrient materials and salts which would change the pH of the soil. Sixteen-inch untreated sticks were selected for uniformity within a very narrow density range and exposed in each of six specially prepared plots in northern New Jersey. The ground was first plowed, then harrowed and raked free of stones so that the soil in all plots was nearly uniform before treatment. Then fifty sticks were buried to a depth of 7 inches in each plot in rows of five, with two feet between each row and one foot between the sticks in each row. The plots were treated as follows:

Plot No.	Treatment
1	Control
2	Barnyard manure
3	5 pounds lime
4	5 pounds commercial fertilizer (5–10–5)
5	5 pounds aluminum sulfate
6	Nutrient solution*

* Containing the following minerals dissolved in ten gallons of water, then sprinkled over the entire plot:

	grams
A) $Ca(NO_3)_2 \cdot 4H_2O$..	297.0
$MgSO_4 \cdot 7H_2O$...	118.8
KH_2PO_4...	83.6
B) $ZnSO_4$..	0.176
$MnSO_4 \cdot 7H_2O$...	0.572
Boric Acid..	0.704
$Al_2(SO_4)_3$...	0.176
C) $FeSO_4$..	5.0

Note: The salts in A, B, and C were dissolved separately and then the three parts mixed.

The sticks were removed and examined at intervals of three, twelve, and fifteen months. The percentage failure, as determined by the ease with which the specimen could be broken, is shown in the following:

	Percentage Failure		
	3 mo.	12 mo.	15 mo.
Control....................................	2	8	72
Manure....................................	10	12	96
Lime.......................................	2	10	80
Fertilizer..................................	4	8	94
Aluminum sulfate.........................	16	36	100
Nutrient solution.........................	4	12	100

At the end of 12 months the greatest number of failures due to decay was observed in the plot treated with the acid salt (aluminum sulfate). Colorimetric determinations of the pH of the soil showed it to be between 5.8 to 6.0 for the soil treated with the acid salt and about 6.6 to 6.8 for the control plot. This experiment needs to be repeated for confirmation of results, but the present indications are that the acid soil was much more favorable for decay than the soil in the control plot. Limed and fertilized soils gave results comparable to those of the control plot, with an indication that the fertilizer increased the decay. The complete disintegration of the sticks in the acid treated soil was particularly noticeable, whereas the sticks in the soil treated with manure were intact, though easily broken. The plot treated with the nutrient solution showed the same rate of decay as the manure plot at the end of the twelve-month period. Ten pounds of aluminum sulfate were then added to the nutrient plot, and three months later all the sticks were completely disintegrated. The rapid disintegration of the sticks in the plot treated with the acid salt points to the importance of further work on the effect of the pH on the rate of decay.

Comparison of Soil Technique With Other Toxicity Assays

In the selection of a method for testing relative toxicity of chemicals to microorganisms, the rapidity of test, the standardization of the medium,

the choice of test organisms, the ease of manipulation, the replication of results, and the duplication by other investigators have been the paramount objectives.

A hypothetical test which would meet these requirements could be performed with distilled water to which could be added increasing concentrations of the chemical to be tested. A known amount of fungus mycelium or spores could be shaken with the toxic solution and left for several different time intervals. The fungus filaments or spores could then be removed to a nutrient agar, and the viability of the fungal filaments or percentage of spores germinating could be readily determined. Comparative toxicity of a large number of compounds could be quickly and easily ascertained. The results, however, would be applicable only to a distilled water-poison system, and the concentration of most toxic materials necessary to inhibit growth would be very low.

The addition of nutrients would necessitate larger amounts of the toxic materials (Van den Berge,[15] 1935). Therefore the mineral solutions— nutrient agar, soil extract agar, soil or wood substrate would in general necessitate an increase in toxic material, the amount of increase depending upon which substrate best meets the nutritional requirements of any particular fungus. Some toxicity values would also be affected by chemical or certain physical changes resulting from interaction between the toxic material and the substrate.

In the petri dish method, the fungi selected for studies of wood destruction grow well on the nutrient substrate containing 1.5% malt extract and 2% agar. Although such a mixture has been recommended as a standard substrate,[2] it should be pointed out that the malt syrup is somewhat variable in composition and constituents and that even the agar varies in the amounts of various growth substances present, Robbins and Ma,[15] 1941. The interpretation of results obtained by the assay of a fungicide when dispersed in an agar system should be restricted to that specific system and not applied to a wood-fungicide system.

When wood preservation studies are carried out, reliance cannot be placed on the results of petri dish tests. The use of wood permits the testing of a large variety of the more common preservatives and fungus-proofing agents, many of which may react with the wood or are precipitated in the wood upon loss of solvent. Organic preservatives which are relatively insoluble in water are not readily tested by petri dish assay.

Comparison of the wood-soil contact method with the wood-water method when untreated wood blocks are used is shown in Fig. 2. The greater uniformity in the amount and rapidity of decay and the better control of moisture showed the soil technique to be superior. It is obvious that if the amount of decay is variable and adversely affected by other factors,

the effect of the preservative or fungicide will be obscured. When the sapwood-water method[3] was published, comparison between it and the kolle flask method showed that the wood-water method had certain advantages.

Comparing the method of soil contact with that of soil burial, the principal point of difference is that a pure culture is used in the soil contact method and a mixed culture is used in the soil burial method. Common to both are the moisture-regulating and nutrient properties of the soil. Since the microbial activity of unsterile soil is diverse, depending on the type and source of soil, uniform results from soil burial could not be expected. When wood specimens were exposed individually in bottles of non-sterile soil in the laboratory, the amount of decay after 12 weeks' exposure was less than 10% for all specimens. The results were similar to those obtained by the exposure of untreated wood out-of-doors at Gulfport, Mississippi, for the two-month period (Fig. 5). Since decay-producing organisms were shown to be present, the other organisms in the soil must have interfered with the growth of the wood-destroying fungi. The antagonism between the wood-destroying fungus *Lentinus lepideus* and a contamination is shown in Fig. 6.

The soil-contact technique instead of the soil burial method has been used extensively to test cotton fabric, thread, paper, jute, fibers, and a variety of other materials. The organisms have been varied according to their occurrence on the particular substrate in nature. The fungi *Chaetomium globosum*, *Aspergillus niger*, *Stachybotrys atra*, *Stysanus media*, and *Metarrhizium* have been established with excellent results on a substrate of cloth when testing fabric. The loss in tensile strength of an unprotected cotton thread which had an initial absolute pull of 30 pounds was 90–100% after two weeks' exposure to *Chaetomium globosum*. Treated threads or other cellulosic materials may be tested as satisfactorily as treated wood.

Examination of numerous reports from soil burial studies of treated textiles indicates that organisms which tolerate certain types of chemicals become dominant in the test beds. As a result, a preservative which shows great promise initially may suddenly fail when the test is repeated. If a pure culture technique were used, a better evaluation of the preservative would be possible.

Similarly the controversies which have arisen over the ability of certain fungi to destroy cellulose could be resolved by using the suitable cellulose soil technique. At least there is very good evidence that many of the environmental variations affecting decay are at or near the optimum.

TOXICITY TESTS

In the toxicity tests which follow, petri dish results are given for several compounds, soil contact test results are given for compounds not readily

assayed by the petri dish method, and the field test results are included for comparison with the results of the soil contact test method.

Fig. 6—Antagonism which has persisted for over one year between the wood-destroying fungus, Lentinus lepideus (outer portion) and contaminating fungus (inner portion) in a soil culture.

During the initial stages of this research on the evaluation of toxic properties of various compounds for wood preservation, the petri dish method was used for assay studies in the laboratory and the modified sapling method

of Waterman and Williams[14] (1934) was used for the field tests at Gulf-
port, Mississippi.   Table 2 shows the results obtained with the petri dish
method on the toxicity of four common inorganic salts and a creosote to
several of the usual test fungi.

TABLE 2

TOXICITY EXPRESSED IN PER CENT TOXIC AGENT PRESENT IN NUTRIENT AGAR AS
DETERMINED BY PETRI DISH ASSAY

Compound	Fungi	Inhibition Point	Killing Point
Arsenic Trioxide	Madison #517	0.04	0.064
	Poria incrassata	0.10	0.10
	Lentinus lepideus	0.30	0.30
	Fomes roseus	0.30	0.30
	Poria microspora	0.30	0.30
	Polyporus vaporarius	0.49	0.49
Zinc Chloride	*Poria incrassata*	0.16	0.16
	Madison #517	0.15	0.23
	Lentinus lepideus	0.16	0.40
	Fomes roseus	0.64	0.64
	Poria microspora	1.40	1.90
	Polyporus vaporarius	1.40	1.90
Mercuric Chloride	Madison #517	0.0012	0.0012
	Lentinus lepideus	0.002	0.002
	Polyporus vaporarius	0.002	0.002
	Poria incrassata	0.005	0.005
	Poria microspora	0.01	0.01
	Fomes roseus	0.01	0.01
Copper Sulfate	*Lentinus lepideus*	0.06	0.16
	Madison #517	0.10	0.16
	Lenzites sepiaria	0.30	0.30
	Fomes roseus	0.24	0.36
	Poria incrassata	0.50	0.50
	Polyporus vaporarius	1.00	1.00
	Poria microspora	1.0	1.00
Creosote	*Poria incrassata*	0.012	0.096
	Polyporus vaporarius	0.024	0.12
	Lenzites sepiaria	0.96	1.00
	Poria microspora	0.20	1.40
	Lentinus lepideus	0.96	1.60

Resistance of the fungi to the four salts is variable, but it will be noted
that *Lentinus lepideus,* which is most sensitive to copper sulfate, tolerates
the highest concentration of creosote.   *Poria incrassata,* which is fairly
tolerant of copper salts by petri dish test, is the most sensitive to zinc chloride
and creosote.   *Poria microspora* tolerates relatively greater concentrations
of all the compounds than any of the fungi tested.

The four salts assayed can be easily dissolved in an agar medium in con-
centrations high enough to be toxic, but uniform dispersal of insoluble salts

in the agar is not so easily accomplished. Many salts may be made soluble
by dissolving them in dilute ammonia or acetic acid solutions. For example,
copper arsenate or zinc meta-arsenite are soluble in ammonia or acetic acid,
and by evaporation of the volatile portions of the solvent the salts are pre-
cipitated. When precipitation of the salts from ammoniacal or acetic acid
solution is carried out in treatments of wood, subsequent evaporation of
ammonia or acetic acid from the wood is rather rapid. In agar solutions,
uniform precipitation of the salts through evaporation of the ammonia and
acetic acid is not easily attained.

TABLE 3

TWENTY FOUR WEEK SOIL ASSAY OF WOOD PRESERVATIVE COMPOUNDS NOT READILY
ASSAYABLE BY PETRI DISH METHODS

Mixture *1	Average Weight Loss in Per Cent			
	1.60 lbs/cu.ft.	0.80 lbs/cu.ft.	0.41 lbs/cu.ft.	Untreated*
Poria incrassata..............	0.0	10.1	10.1	61.8
Polyporus vaporarius..........	0.0	0.7	0.6	34.1
B.T.L. U-11................	0.0	1.8	6.4	56.3
Mixture *2	2.79 lbs/cu.ft.	1.40 lbs/cu.ft.	0.68 lbs/cu.ft.	
Poria incrassata..............	40.9	31.9	38.8	52.0
Polyporus vaporarius..........	45.9	30.6		48.6
B.T.L. U-11................	21.3	57.0	31.8	48.5
Mixture *3	0.35 lbs/cu.ft.	0.15 lbs/cu.ft.		
Poria incrassata..............	46.7	22.9		52.4
Polyporus vaporarius..........	0.0	6.1		63.8
B.T.L. U-11................	0.0	5.9		56.4
Mixture *4	0.72 lbs/cu.ft.	0.36 lbs/cu.ft.		
Poria incrassata..............	1.0	13.8		50.7
Polyporus vaporarius..........	0.0			55.4
B.T.L. U-11................	1.0	14.5		53.6

* Average per cent weight loss of untreated blocks in the same bottles with the treated
blocks.

Agar cannot readily be used for assays of two other types of compounds
used as wood preservatives. The first type depends on chemical reactions
with and also within the wood. Specific examples of this type are the series
of compounds fixed in the wood by the reduction of chromium salts which was
first studied by Kamesam,[17] 1934. It is now the generally accepted view
that the reduction of the chromium is brought about by various sugars in
the wood. Subsequent research led to the use of Ascu (Kamesam) or Green-
salt K and to the later development of Greensalt "O" by the Bell Telephone
Laboratories in the United States and the Bolidens' salts in Sweden. These
inorganic salt mixtures were developed in the search for preservatives which

would be fixed in the wood and thus resist leaching when exposed to the action of ground waters.

The second type is comprised of organic compounds or mixtures of organic compounds, such as creosote, which has had an excellent service record as a preservative. Also included in this type of compounds (which are insoluble in water and have a relatively low vapor pressure) are certain chlorinated phenols and cresols. Because of the low water solubility or immiscibility with agar solutions, the uniform dispersal of the toxic agents in the agar system, which is essential to reproducibility of results, is almost impossible. Uniform injection of these materials into wood, however, presents no particular problem.

Assays of four representative mixtures not readily assayable in the petri dish are included in the results of Table 3. The composition of the treating solutions of the mixtures 1, 2, 3, and 4 is given below:

Mixture 1	Zinc oxide	46.6%
	Chromic acid	3.8%
	Arsenic acid	49.6% dissolved in a 10% ammonia solution
Mixture 2	Copper phenolate	4%
	Zinc phenolates	1%
Mixture 3	Sodium fluoride	25%
	Disodium hydrogen arsenate	25%
	Sodium chromate	37.5%
	Dinitrophenol	12.5%
Mixture 4	Zinc oxide	22.5%
	Arsenic oxide	35.5%
	Sodium carbonate	1.0%
	Acetic acid	41.0%

In the assays shown in Table 3 the maximum retention of the mixture by the wood is that recommended for the treatment of wood that is not to be used in contact with the ground. To make the test as severe as possible, organisms were selected which were known from petri dish, kolle flask, and wood-water assays to have a high tolerance for various inorganic salts.

Examination of the data in Table 3 shows that the compounds produced in the wood by mixture 4 afford almost complete protection to the wood which was treated with 0.72 pound of the salt per cubic foot. The vigorous attack on the untreated blocks is evidence of the severity of the test. Similarly, the comparable treatment of the wood with 0.8 pound of mixture 1 per cubic foot was very effective in protecting the wood against fungus attack. If the amount of mixture 1 in the wood is doubled, almost perfect protection against decay may be obtained. Wood treated with 0.35 pound of mixture 3 per cubic foot is protected against *Polyporus vaporarius* and BTL U-11 but

not against *Poria incrassata.* The latter fungus completely disintegrates both the treated and the untreated wood. Despite the fact that the wood treated with mixture 2 represented the highest concentration of any preservative used (2.8 pounds per cubic foot), complete disintegration of the wood results from the action of the three fungi just mentioned.

Compounds of the Greensalt type were also assayed by means of the soil-contact method. The solution commonly used for treatment of wood with Greensalt K contains three chemicals in the following proportions:

Potassium dichromate	$K_2Cr_2O_7$	55%
Copper sulfate	$CuSO_4 \cdot 5H_2O$	33%
Arsenic acid	$As_2O_5 \cdot 2H_2O$	11%

After treatment of the wood with this solution, reduction of the chromium by the sugars in the wood together with evaporation of water precipitates in the wood fibers several complex insoluble salts, among which presumably
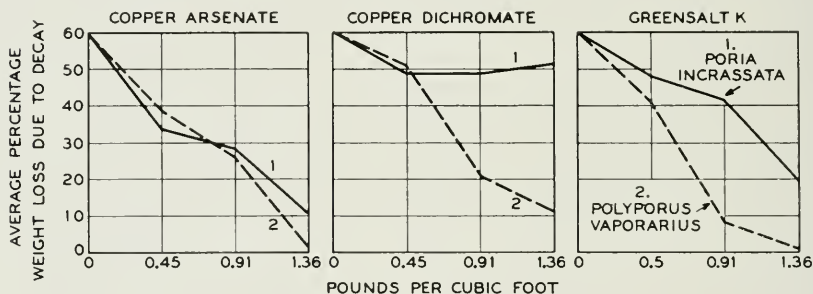
Fig. 7—Comparison by wood-soil assay of components with the whole salt of Greensalt K. Organisms Poria incrassata and Polyporous vaporarius. 24 weeks' time.

are copper arsenate and copper dichromate. Results of soil-contact assay of wood treated separately with solutions of these two components and with the whole Greensalt K complex are given in Fig. 7. Before exposure to the fungi, the wood specimens were leached by a diffusion method described by Waterman, Leutritz and Hill[3], 1938. No untreated control blocks were included in the bottles with this test, which was conducted for 24 weeks. The copper arsenate component of the Greensalt K complex is shown to be much more effective as a preservative than the copper dichromate component but not as effective as the whole K salt complex. Figure 8 shows the setup for the wood-soil assay of 0.75 lbs/cu. ft. of Greensalt K from a recent series of cooperative experiments conducted by the Forest Products Laboratory, Madison, Wisconsin. Figure 9 is a comparison of the Greensalt K treated and untreated reference blocks in the same bottles after exposure to the following fungi:

Blocks	Fungi
A	*Lenzites trabea* ✻ 617 F.P.L.
B	*Poria incrassata* ✻ 563 F.P.L.
D	*Lentinus lepideus* ✻ 534 F.P.L.
E	*Poria microspora* ✻ 106 F.P L.
F	*Poria luteofibrata* (Baxter)

The distortion and shrinkage of the blocks can be used as a visual confirmation of the weight loss due to decay.

The more recently developed Greensalt O is similar to the K salt. The treating solution of this salt mixture is composed of copper oxide, hydroxide or carbonate, chromic acid anhydride, and arsenic acid in percentages based on the chemical equivalents of the copper, chromate and arsenic salts in the K salt solution. The toxicity from the wood-soil assay of Greensalt O is given in Table 4 for fourteen fungi and for three concentrations of the preservative. The effectiveness of the Greensalt O treatment of wood is apparent from examination of the toxicity index. The weight losses due to decay of the untreated reference blocks indicated in general that conditions for decay were again very severe, but the reference blocks exposed to the fungus *Lentinus lepideus* were protected by their proximity to the treated specimens. As previously pointed out, the fungus *Lentinus lepideus* does not tolerate even slight concentrations of copper, which is a major component of the Greensalt complex.

The fungus *Poria incrassata* was again shown to be the least affected by the toxicity of the preservative. The toxicity index for the blocks treated with the highest concentration of the preservative was 94% after 16 weeks' exposure to this organism. In view of the excellent field record for the equivalent K salt preservatives ratings 90 to 100% by the toxicity index would be satisfactory. Additional data will undoubtedly determine the limits of the toxicity index.

At the end of the 24-week period, the fungi BTL U-11, *Lenzites trabea*, *Trametes serialis*, *Polyporus vaporarius*, and one strain of *Coniophora cerebella* caused slight decay of one or more blocks treated with the maximum concentration of preservative, the fungi *Lenzites trabea*, *Coniophora cerebella*, and *Trametes serialis* were still capable of causing only slight losses. Exposure of the blocks treated with the low contentration of preservative to *Polyporus vaporarius* and BTL U-11 resulted in an increase in the amount of decay.

Since the resultant salts of the Greensalt O reaction should be similar to those produced by Greensalt K, field trials of these materials would be expected to give comparable results. Extended field tests of wood treated with one pound of Greensalt K per cubic foot have been in progress for ten years without a single failure having occurred in more than 40 specimens. Specimens treated with Greensalt O have been tested in the field for only

Fig. 8—Wood-soil assay of 0.75 lbs/cu. ft. of Greensalt K (courtesy of Forest Products Laboratory, Madison, Wisconsin). Reading left to right the samples are:

Blocks   Fungi
A—Lenzites trabea #617 F.P.L.
B—Poria incrassata #563 F.P.L.
D—Lentinus lepideus #534 F.P.L.
E—Poria microspora #106 F.P.L.
F—Poria luteofibrata (Baxter)
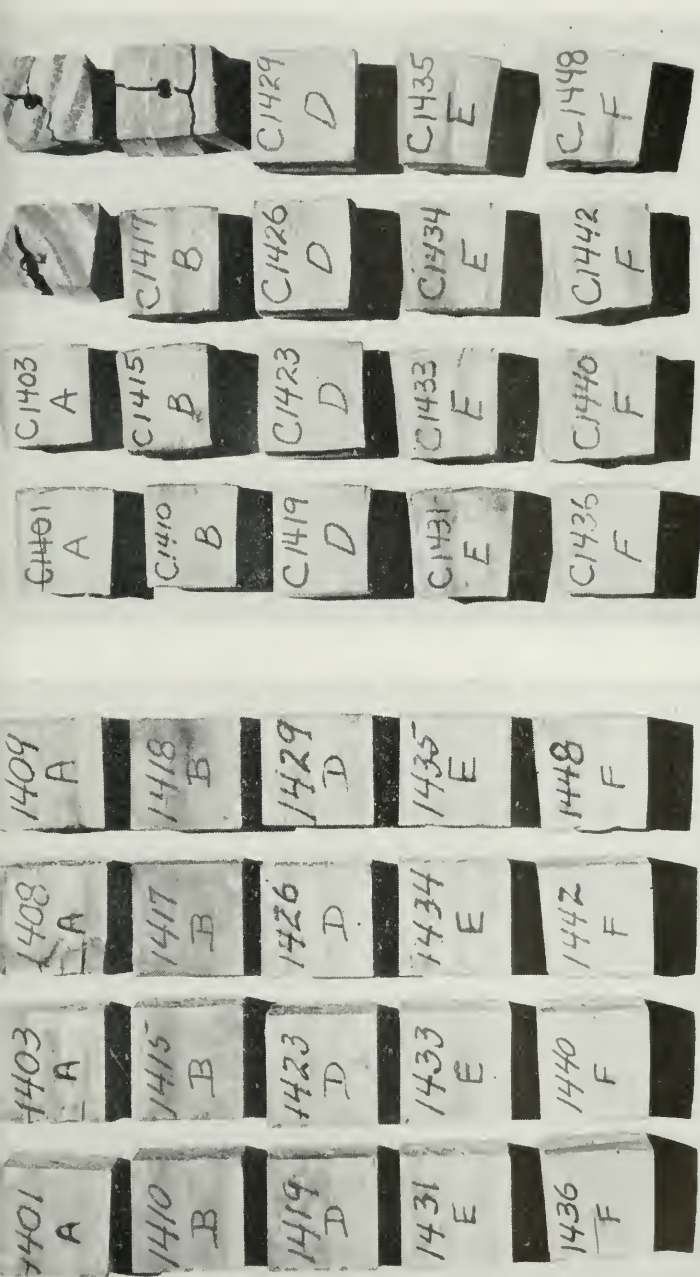
Fig. 9—A comparison of the Greensalt K treated and untreated reference blocks after exposure to fungi.   (Treated on the left—untreated on the right side of picture.)

Blocks

Fungi

A—*Lenzites trabea* #617 F.P.L.
B—*Poria incrassata* #563 F.P.L.
D—*Lentinus lepideus* #534 F.P.L.
E—*Poria microspora* #106 F.P.L.
F—*Poria luteofibrata* (Baxter)

the relatively short period of three years, during which time all the specimens have remained sound.

Results of other field trials for a three-year period with the same mixtures 1, 2, 3, and 4, listed previously, copper arsenate, Greensalt O, and a creosote are given in Fig. 10. Twenty-five untreated controls showed 84% failure, 4% sound, and 12% badly infected in one year. All had failed at the end of the second year. Twenty specimens were used for each retention of the individual preservatives, with the exception of the creosote. The reason for fewer creosote specimens within the correct retention is that the empty-cell treatments of wood with creosote give a wider range of retention than the full-cell treatments of the wood with water solutions of the salts.

TABLE 4

TWENTY FOUR WEEK—WOOD SOIL ASSAY OF GREENSALT "0" USING 13 SPECIES OF WOOD DESTROYING FUNGI

	Toxicity Index*		
	1.17 lbs/cu.ft.	0.96 lbs/cu.ft.	0.476 lbs/cu.ft.
Poria incrassata (16 wks)	94	68	25
B.T.L. U-11	98	88	62
Polyporus vaporarius	98	93	55
Lenzites trabea	98	96	94
Coniophora cerebella	98	98	98
Trametes serialis	98	98	98
B.T.L. U-4	100	98	98
Polyporus anceps	100	98	98
B.T.L. U-53	100	99	98
B.T.L. U-24	100	99	98
Lenzites sepiaria	100	100	98
Poria microspora	100	100	98
Fomes roseus	100	100	100
Lentinus lepideus	100	100	100

$$* \text{ Toxicity Index} = \frac{\% \text{ loss of norm} - \% \text{ loss of treated block}}{\% \text{ loss of norm}} \times 100.$$

In the three-year period of exposure only the treatments of wood with 1.3 pounds of copper arsenate and 7 pounds of creosote per cubic foot showed a perfect record. Copper arsenate had previously been shown to be a very effective component of the Greensalt complexes when tested by the soil-contact method. Mixture 2 was found to be the poorest by soil-contact assay and also in the field trials. Mixtures 4, 1, and 3 were rated in that order of decreasing effectiveness in the field test. In the soil-contact assays at comparable retention, 0.72 and 0.80 pound of salt per cubic foot of wood, respectively, mixture 4 was better than 1; and at 0.41 and 0.35 pound of salt per cubic foot of wood, respectively, compound 1 was slightly better than 3, especially against *Poria incrassata*.

Results from field and laboratory tests show good agreement in the evalu-

ation of the compounds. The advantage of the laboratory method in the matter of time is a decided one, but the field trial is valuable for testing the permanence of the preservative. For example, mixture 4 was found in the laboratory test to be a very effective preservative, but the initial preservative properties were dissipated by exposure to the weather, since 50% of the
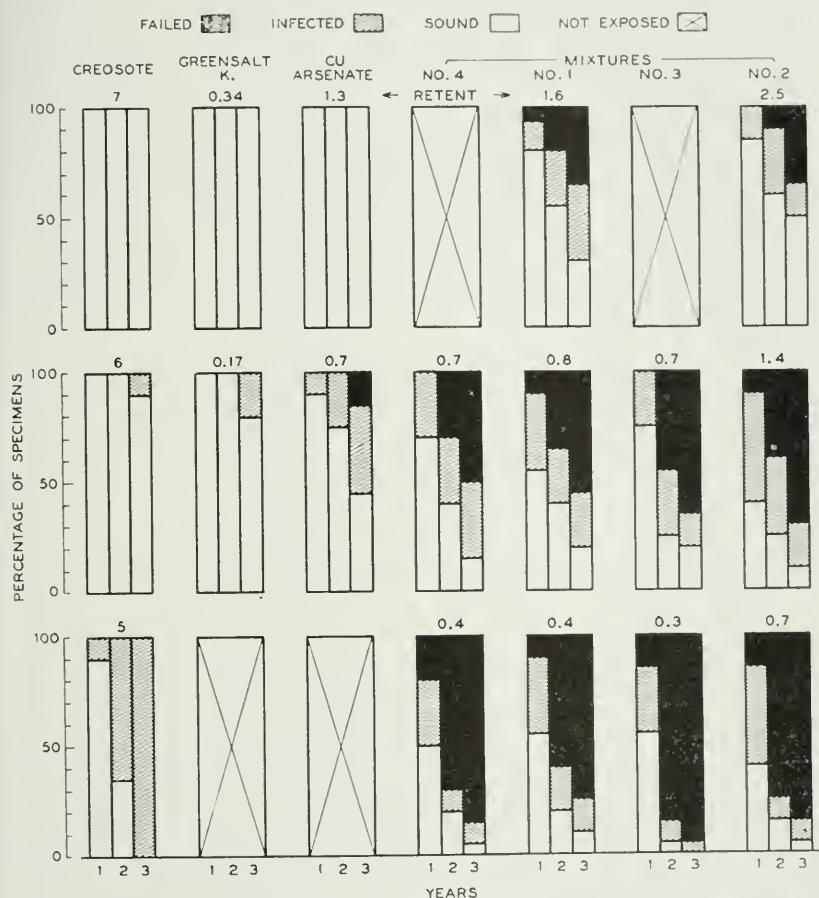
Fig. 10—Results of field exposures on the four mixtures of Table 3 and other compounds creosote, Greensalt K, copper arsenate (retent in pounds per cubic foot).

specimens had to be removed because of failure within three years. When a compound is a poor preservative, as in the case of mixture 2, both laboratory and field trials serve to eliminate it from further consideration. By the soil-contact method every specimen treated with mixture 2 was badly attacked by all the organisms used, whereas at a comparable retention only 35% of the specimens in the field trials had failed after three years' exposure.

Since these results indicate that the soil-contact provides optimum conditions for decay, the laboratory method serves as a means of quickly eliminating inferior preservatives and minimizing the number for field studies.

Summary

The soil-contact method described in this paper has been shown to be a valuable laboratory tool for the study of fungus destruction of cellulose and wood and for the determination of the value of wood and cellulose preservatives.

Top soil containing 20 to 25% moisture on a dry-weight basis, when used as a supporting substrate for decaying wood, proved to be an excellent means controlling the moisture content of wood during the decay process. Investigation showed that the optimum moisture content for initiating the decay of wood was fiber saturation. It also was found that during decay the initial water content of the wood remained constant, through maintenance of a constant volume of the wood structure despite loss of wood substance.

Experiments with various combinations of nutrients and nutrilites added to artificial soil showed the importance of these materials in decay studies. The need for nitrogen in the destruction of cellulose by fungi was confirmed. Lack of wood decay in the presence of a sugar when there is also a deficiency of nitrogen presents an interesting problem the explanation of which may throw considerable light on discrepancies in many test procedures. Comparison of results with nutrient artificial soils and an average top soil indicates the possibility of employing a standard artificial soil in the contact test method.

The optimum temperature for most wood-destroying fungi tested was found to be 26°–28°C. Decay occurred over a wider range of temperature in soil-contact tests than in petri dish tests.

It was found that decay was much more uniform and more rapid in the soil-contact method than in other laboratory methods or in field trials. There is a large, single, vigorous inoculum in the soil-contact laboratory method, while in the field antagonism between wood-destroying organisms and the other flora and fauna of the soil frequently checks the decay process.

Toxicity studies based on petri dish assays showed that the amount of a compound tolerated by several fungi varies considerably. Petri dish assays of toxic materials are often misleading. Generally, higher retentions of the preservatives are needed to prevent decay than are indicated by petri dish assay. Occasionally, a material which performs poorly in the petri dish test will, however, act as a satisfactory preservative of cellulosic derivatives in both soil-contact and field tests.

Field trials of preservatives, though in general less rapid, confirm the results of the soil-contact method and in addition determine the degree of

permanence of the preservative. However, heating, leaching and other simulated weathering cycles may be used in conjunction with the soil-contact method to determine the stability of a preservative to evaporation, to the solvent action of ground water, and to chemical deterioration by ultra-violet light. Futher comparisons between soil-contact assay and field test of a wood preservative such as Greensalt confirms the fact that conditions for decay are more nearly optimum in the former and result in an unusually severe test of the preservatives.

The soil-contact method is an excellent laboratory tool easily adapted to fundamental studies of the decay process and to evaluation of preservatives. The method has shown considerable promise in evaluating preservatives for a wide variety of materials, including leather, cotton, felt, paper and jute. Since the factors influencing decay are very near optimum in the soil-contact method, any preservative that prevents decay in this laboratory test and is permanently retained will be effective under any climatic conditions.

ACKNOWLEDGMENT

BIBLIOGRAPHY

1. Williams, R. J., 1928, "Nutrilites," *Science 67,* 607.
2. Richards, C. A., 1923, "Methods of testing relative toxicity," *Proc. Amer. Wood-Pres. Assoc., 19.*
3. Waterman, R. E., Leutritz, John, and Hill, C. M., 1937, "A Laboratory Evaluation of Wood Preservatives, *Bell Sys. Tech. Jour., 16,* 194–211.
4. Falck, R., 1927, "Merkblätter zur Holz-Schutzfrage," *Hausschwammforschungen begrundet, von A Möller, Heft 8,* p. 17, *Jena.*
5. *American Society for Testing Materials,* 1942, "Tentative Methods of Test for Resistance of Textile Fabrics to Microorganisms," *A.S.T.M. Designation:* D684–42T.
6. *Signal Corps Tentative Specification* No. 71-2202A, April 12, 1944, "Moisture and Fungus-Resistant Treatment of Signal Corps Ground Signal Equipment."
7. Furry, Margaret S. and Zametkin, Marion, 1943, "Soil-suspension Method for Testing Mildew Resistance of Treated Fabrics," *AM Dyestuffs Reptr., 32,* 395–8.
8. Leutritz, John, Jr., 1939. "Acceleration of Toximetric Tests of Wood Preservatives by Use of Soil as a Medium," *Phytopathology, 29* (10) 901–903.
9. Nobles, Mildred K., 1943, "A Contribution Toward a Clarification of the *Trametes Serialis* Complex." *Canadian Jour. of Res., 21,* 211–234.
10. Schorger, A. W., 1926, "The Chemistry of Cellulose and Wood," McGraw, New York.
11. Benton, V. L. and Ehrlich, John, 1941, "Variation in Culture of Several Isolates of *Armillaria Mellea* from Western White Pine," *Phytopathology, 31,* (9) 803–811.
12. Schmitz, Henry and Kaufert, Frank, 1936, "The Effect of Certain Nitrogenous Compounds on the Rate of Decay of Wood," *Amer. Jour. of Botany, 23,* (10) 635–8.
13. Humphrey, C. J. and Siggers, P. V., 1933, "Temperature Relations of Wood-Destroying Fungi," *Jour. of Agr. Res., 47,* (12) 997–1008.
14. Waterman, R. E. and Williams, R. R., 1934, "Small Sapling Method of Evaluating Wood Preservatives," *Ind. and Engr. Chem., Analytical Ed., 6,* 413–419.
15. Van den Berge, J., 1935, "Testing and Suitability of Fungicides for Wood Preservatives." Unpublished manuscript. Translation from the Dutch.
16. Robbins, W. J. and Ma, Roberta, 1941, "Biotin and the Growth of *Fusarium Avenaceum, Bull. Torrey Bot. Club,* 68 (7) 446–462.
17. Kamesam, Sonti, 1934, *British Patent* 404,855.

X-Ray Studies of Surface Layers of Crystals

By ELIZABETH J. ARMSTRONG

1. INTRODUCTION

WHEN a crystalline substance is sawed, ground, lapped or polished, the crystal structure adjacent to the worked surface is distorted and ruptured. Since the selective diffraction of X-rays by a crystal is a result of the orderly arrangement of the planes of atoms of the crystal, disturbance of this arrangement is detectable by X-ray diffraction.

Research on aging of quartz oscillator plates seems to indicate that changes which are accelerated by high humidity take place in this disturbed material resulting in changes in the frequeny and activity[1] of the crystal plate. A knowledge of the nature and extent of this disturbed layer is essential to an understanding of the changes that are taking place in it and may contribute to the improvement of the quality of crystal plates, apart from the problem of aging.

The purpose of this paper is to present a review of X-ray techniques that have been or may be useful tools for the examination of the nature of the surface layers of crystals. Each technique is also discussed from the standpoint of the kind of evidence which it seem best suited to bring to light. Familiarity with the general principles of X-ray diffraction as outlined in this Journal, volume xxii, number 3, pages 293 and 297, is assumed.

It has long been known that the nature of the surface preparation of a crystalline substance affects the intensity of the reflected X-rays. As early as 1913, about a year after the first X-ray diffraction experiments with crystals, de Broglié and Lindemann[2] noticed that the spots in Laue photographs of certain crystals were inhomogeneous and suggested the interpretation that the darker parts of the spots might result from disturbed material. A. H. Compton[3], using a double crystal spectrometer in 1917, found that the reflection from a ground surface of a calcite crystal was three times that from a cleavage face.

---

[1] One plate is said to have greater activity than another similar plate if its amplitude of oscillation is greater when the two are tested under identical conditions. The activity of a plate is reduced by friction with its mountings or with particles on its surface, either of quartz or of a foreign material.

[2] de Broglié, M. and Lindemann, F.-A., "*Sur les Phénomènes Optiques Présentés par les Rayons de Röntgen Rencontrant des Milieux Cristallins, Comptes Rendus,* 156 (1913), pp. 1461–1463.

[3] Compton, A. H., "*The Reflection Coefficient of Monochromatic X-Rays from Rock Salt and Calcite,*" *Phys. Rev.,* 10 (1917), pp. 95–96.

The explanation of the higher intensity of reflection from the disturbed surface lies in the fact that the rays of the incident X-ray beam, collimated by a pair of slits, are not perfectly parallel, but diverge, meeting the crystal plate at various angles, whose range, depending on the geometry of the slits, is usually about 15 to 25 minutes of arc. If the surface regiln of the crystal plate is undisturbed only a very small part of the incident beam will meet the crystal at the Bragg angle for X-ray reflection. If, however, some of the quartz has been disturbed it will have a variety of orientations with respect to the main crystal structure and the various disturbed bits of quartz will be at the Bragg angle for the various divergent rays of the incident X-ray beam. In this way more of the incident beam is reflected by a disturbed crystal surface than by an undisturbed crystal surface. The disturbed material measured by this technique differs in orientation from that of the main plate by not more than a few minutes so that even this disturbed material uses only a small sector of the divergent incident beam. Surface particles misoriented by larger angles are not numerous enough to reflect X-rays into the ionization chamber with measurable intensity.

An alternative interpretation of the higher intensity of reflection from the disturbed material should be mentioned although it has little practical significance. Consideration of the Bragg equation, $n\lambda = 2d \sin\theta$, will show that a range of d values would make it possible for a range of θ values to satisfy the equation. If, therefore, there were some variability in the spacing, d, between the atomic planes from which the X-rays were being reflected, reflection would take place over a corresponding range of angles of incidence.[4] Such variability in d spacing would be a result of lattice distortion. It would generally be accompanied by misorientation and therefore its consideration as a phenomenon distinct from misorientation becomes rather academic. The disturbance will therefore be spoken of as misorientation although it probably also involves small changes in d spacing.

Measurable lattice distortion can be produced by other means than surface working. The reflection-intensity of an etched plate is increased three or four times if the plate is strained by bending during the reflection of the X-ray beam. When the pressure on the plate is released the reflection-intensity resumes its former value. The distortion produced by unequal pressures on the plate results in the heterogeneity of orientation which makes possible the use of a larger part of the incident beam, resulting in higher reflection-intensity. When lapped plates are similarly deformed the increase in reflection-intensity is less since some heterogeneity of orientation already exists. As would be expected, the effect of the deformation is progressively less with

---

[4] Consideration of the known compressibility and tensile strength of quartz indicates that the maximum change in d spacing which could be obtained would be of the order of 0.1%. For small values of θ the change in θ for this d change would also be 0.1%, increasing, with larger θ values to about 0.2% at $\theta = 70°$.

increasing grain size of the abrasive with which the surface of the plate was
lapped.

2.1   THE SINGLE CRYSTAL SPECTROMETER

The single crystal spectrometer is used for X-ray measurement of
the orientation of quartz oscillator plates.   In this instrument slit-collimated
X-rays are reflected from a crystal into an ionization chamber and the rela-
tive intensity of the reflected X-rays is read from a meter showing the
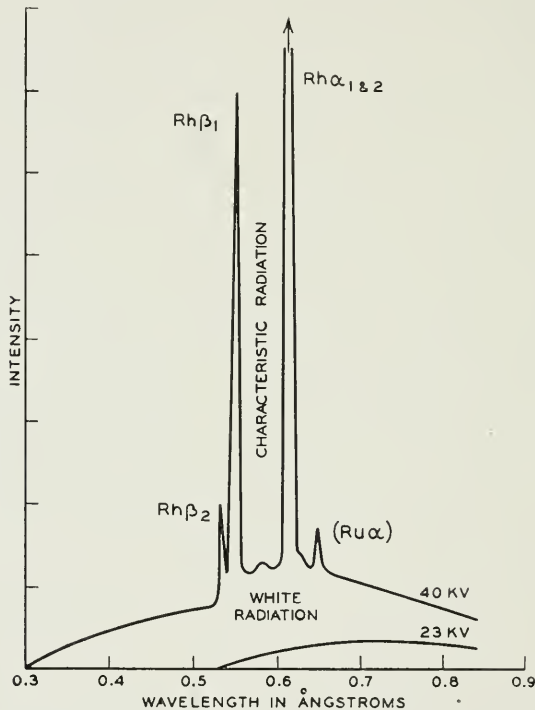
Fig. 1.—X-Ray spectra from a rhodium target at 23 and 40 kilovolts.   Ruthenium im-
purity present.   (Adapted from *Siegbahn, Spectroscopy of X-Rays*).

amplified current resulting from the ionization of the gas in the chamber by
the X-rays.   Since the reflected white radiation is too weak to cause a meas-
urable amount of ionization, only the reflected characteristic radiation is
measured by the ionization chamber.   For most purposes a copper-target
tube is used and the β radiation (comparable to Rh_β of Fig. 1) is eliminated by
a selective filter so that only the α radiation is used, the X-rays thus being
essentially "monochromatic".

Three different techniques for examining surface layers of crystals with
the single crystal spectrometer will be described.   Two of these employ

photographic films or plates in addition to the ionization chamber for measuring the reflected rays.

2.2 Reflection-Intensity Measurements on the Single Crystal Spectrometer with the Ionization Chamber (Fig. 2)

Working with the single-crystal spectrometer, Bragg, James and Bosanquet[5] found the reflection-intensity from a ground face of rock salt two to four times that from a cleavage face. Dickinson and Goodhue[6] found the reflection-intensity from ground faces of sodium chlorate and sodium
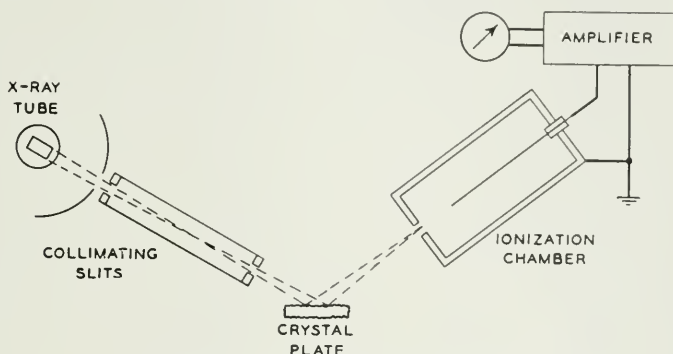
Fig. 2.—Single crystal spectrometer

bromate twice that from the natural face. Sakisaka[7] found the reflection-intensity from a quartz plate lapped with ✳30 carborundum $2\frac{1}{4}$ times that from a plate lapped with very fine emery, which, in turn, was twice that from an etched plate. Recent measurements by the writer have shown that the reflection-intensity of an etched surface increases with progressive lapping and that of a ground surface decreases with progressive etching as shown in Fig. 3. In these figures the reflection-intensity is given in terms of the ratio of the intensity from the test plate to that from a standard plate of the same cut, etched 20 minutes following fine lapping. The initial rate of increase of intensity-ratio with lapping or decrease with etching is very high.

This technique would be most useful in sample-testing groups of plates to check whether they had been inadequately etched or whether any lapping at all had occurred after etching. (In either case the plate would be subject to

[5] Bragg, W. L.; James, R. W. and Bosanquet, C. H.; "The Intensity of Reflexion of X-Rays by Rock Salt," Part I. *Phil. Mag.* 41 (1921) pp. 309–337; Part II, *Phil. Mag.* 42 (1921) pp. 1–17.

[6] Dickinson, R. G. and Goodhue, E. A.; "The Crystal Structure of Sodium Chlorate and Sodium Bromate," *Jour. Amer. Chem. Soc.* 43 (1921) pp. 2045–2055.

[7] Sakisaka, Y.; "The Effects of the Surface Conditions on the Intensity of Reflexion of X-Rays by Quartz," *Japanese Jour. Phys.* 4 (1927) pp. 171–181.

aging.)   Although a plate lapped with 180 carborundum can be easily distinguished from one lapped with $303\frac{1}{2}$ emery by a comparison of intensity ratios, plates lapped with nearly similar abrasives cannot be distinguished with certainty, except on a statistical basis.

That the disturbed material measured by this technique differs in orientation from that of the main plate by less than a few minutes is shown by the fact that the range of incident angles over which ionization-detectable X-ray reflection takes place from a lapped crystal plate is the same within the
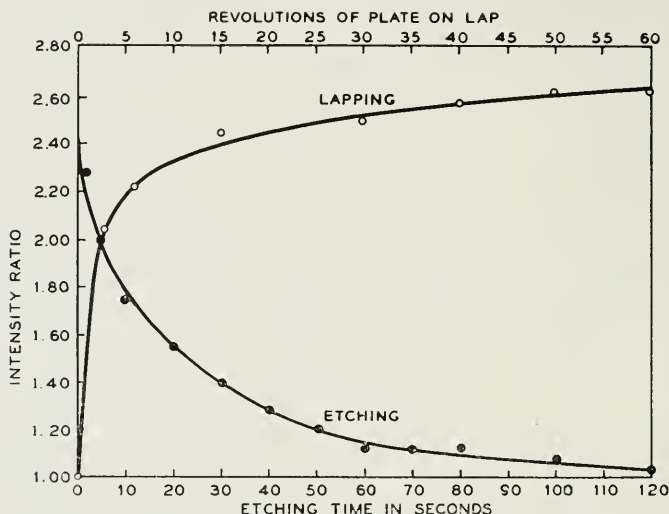
Fig. 3.—Effect on reflection-intensity produced by lapping on etched plate with $303\frac{1}{2}$ emery and soap solution and by etching the lapped plate with 48% hydrofluoric acid at 25°C.

limits of error of measurement as that for reflection from an etched crystal plate.

2.3 PHOTOGRAPHIC MEASUREMENT OF ANGULAR MISORIENTATION WITH THE SINGLE CRYSTAL SPECTROMETER

A technique which does indicate quartz misoriented by more than a few minutes has been devised by Dr. C. J. Davisson. Although the X-rays reflected from this material are too weak to produce a measurable current in the ionization chamber, they will darken a photographic plate or film if adequate exposure time is allowed. The principles of this technique are illustrated in Fig. 4 and some of the resulting photographs are shown in Figs. 5 and 6.

The plate to be measured is placed at the Bragg angle to the incident X-rays as determined by preliminary measurement of the maximum ionization

current produced in the ionization chamber which is at twice the Bragg angle to the incident beam. A film in a paper envelope is then placed before the ionization chamber in a holder which permits a small portion of it to be exposed at a time. A brief (5 or 10-second) exposure is then made with the crystal plate in this "zero position",[8] (see Figs. 4 and 5), recording the strong characteristic radiation reflected from the main crystal plate. At the same
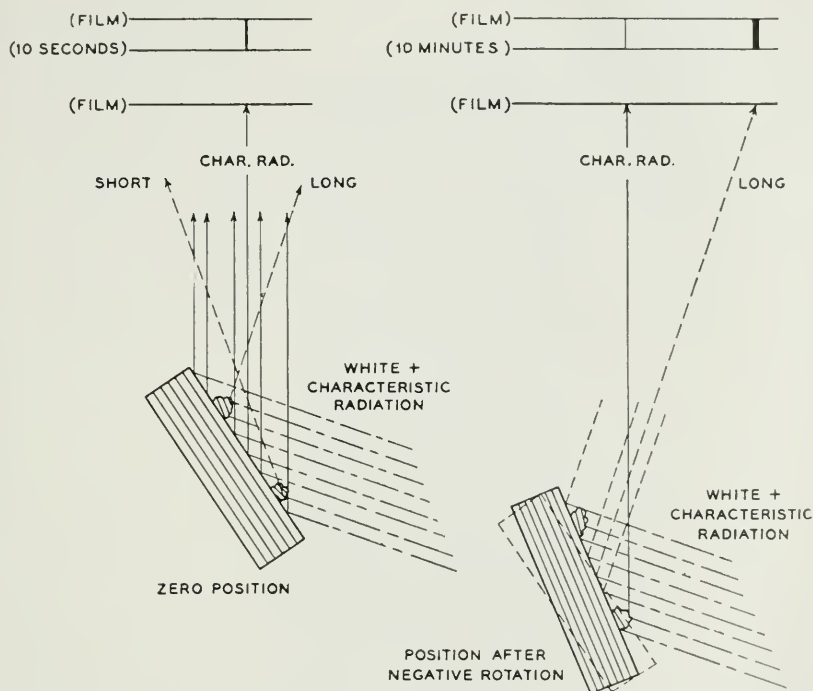
Fig. 4.—Spectrometer photography of misoriented crystalline material

time the weak white radiation is being reflected from the disturbed material, but this radiation is relatively so weak and the disturbed material of such

[8] To check the correctness of the "zero" setting, a "rocking" exposure is taken, during which the plate is rocked through the Bragg angle. The upper half of the beam should be shielded for the "zero" exposure and the lower half for the "rocking" exposure so that the film need not be moved between the two exposures. In the "rocking" exposure the beam is reflected during only a fraction of the exposure time and because the exposure is so brief only the reflection of the strongest part of the incident beam (the part that is going to produce the reflections in later exposures) is recorded. In the "zero" exposure the crystal plate may have been set so as to reflect the divergent, weaker rays of the beam which may differ in direction from the strong part of the beam by as much as 15 minutes. The terms "stronger" and "weaker" do not refer here to characteristic and white radiation, but to parts of the beam that have more or fewer X-rays due to the geometry of the collimating system with relation to the target.

relatively small volume that the reflection does not noticeably darken the photographic film in five seconds. Successive ten-minute exposures are then taken with the crystal plate turned at successively greater angles from the zero position. At any given angle the disturbed quartz that has thus been brought into the proper position to reflect the characteristic radiation does so, producing a center line on the film whose intensity is roughly proportional to
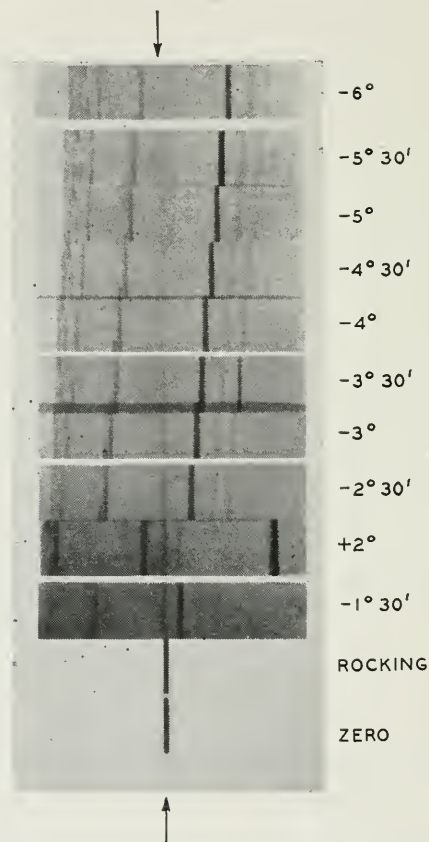
Fig. 5.—Spectrometer photograph of BT quartz plate lapped with 303½ emery

the volume of quartz misoriented to this angle. Various wave-lengths of the white radiation will satisfy the Bragg equations for various atomic planes of the main plate at each angular position and will be reflected to other positions on the film. Although the incident white radiation is relatively weak the reflected beams are strong enough to darken the film in ten minutes because rays reflected from the main plate are reflected by a much greater

volume of quartz than are rays reflected from the disturbed layer. The strongest of these "white" reflections from the main plate is that from the set of atomic planes most nearly parallel to the surface of the plate, the planes that reflected the characteristic radiation in the zero position.

NO.3, 10 SECONDS

−3°
−2° 30′
−2°
−1° 30′
−1°
−30′
ROCKING
ZERO

NO.3, 20 SECONDS

−2° 30′
−2°
−1° 30′
−1°
−30′
ROCKING
ZERO

NO.3, 40 SECONDS

−2°
−1° 30′
−1° 30′
−1°
−30′
ROCKING
ZERO
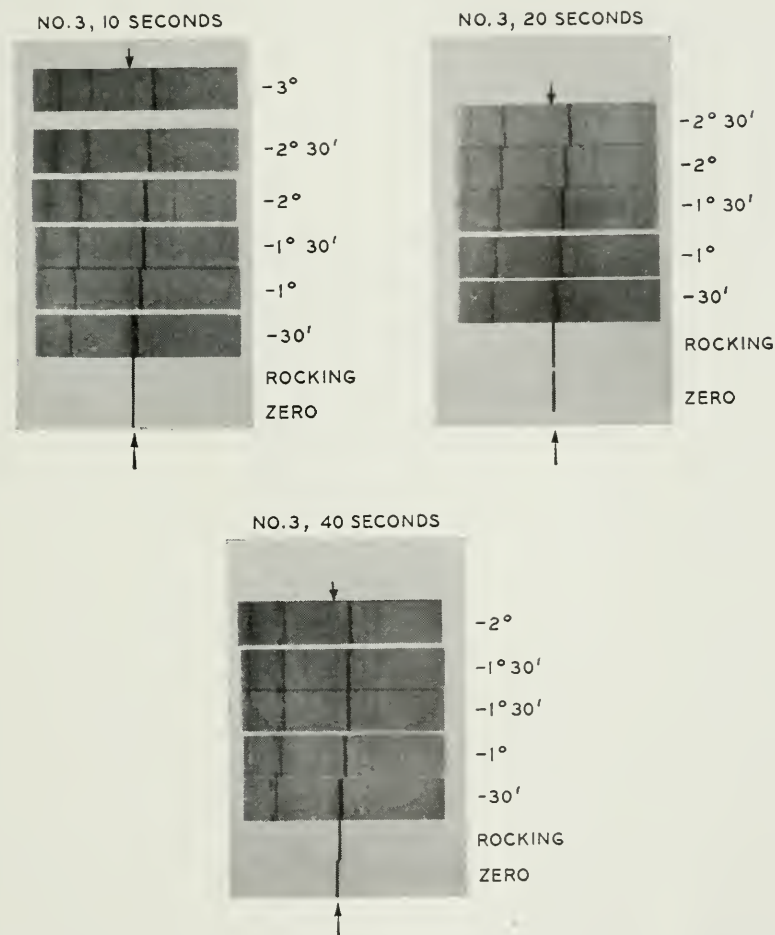
Fig. 6.—Spectrometer photographs showing effect of etching a $303\frac{1}{2}$ emery-lapped BT quartz plate with 48% hydrofluoric acid

In Fig. 5, the central line (indicated by the arrow) resulting from the characteristic radiation reflecting from the disturbed quartz, is distinctly present through the 4° exposure but not in higher-angle exposures, indicating that in the $303\frac{1}{2}$ emery-lapped surface from which the X-rays were reflected there was not enough quartz misaligned by more than 4° to reflect a beam that

would visibly affect a photographic film during a ten-minute exposure.   The
dark line that moves to the right as the negative angular rotation increases
results from the reflection of progressively longer wave-lengths from the
quartz of the main plate.   The three series of exposures in Fig. 6 show the
progressive removal of the disturbed quartz by etching with 48% hy-
drofluoric acid.   After ten seconds' etching the line from the disturbed
material does not show distinctly beyond the 1° 30' position; after 20 seconds'
etching it is distinct only through the 1° 00' position and after 40 seconds it
can only be seen distinctly at the 30' position.   If the film had been ex-
posed for a longer time at each position the line from the disturbed material
at each angle would have persisted with longer etching.   With the arbitrarily
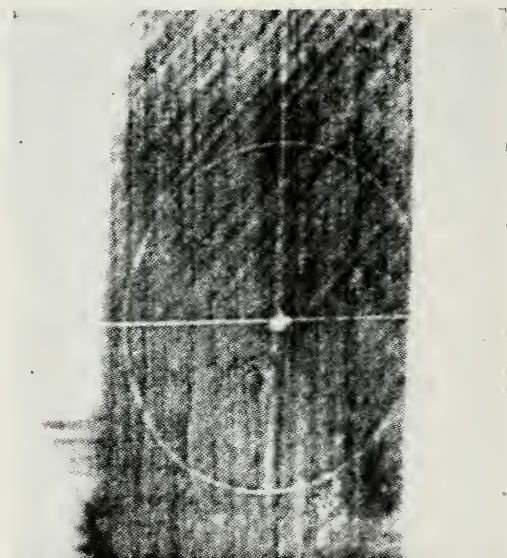
Fig. 7.—Photograph produced by reflection of a broad X-ray beam from the (100)
cleavage face of a rock salt crystal (Berg)

chosen ten-minute exposures a line from material misoriented by 45' dis-
appears after about 25 seconds' etching with 48% hydrofluoric acid, but the
disappearance of the 30' line occurs only after 70 seconds' etching, which
removes an amount of quartz equivalent in weight to a layer about four-
tenths of a micron thick on each surface.

With this technique we are measuring the more grossly misoriented sur-
face material, material that is probably not continuous with the quartz of
the main plate.   This is evident from the fact that a piece of quartz would
have to have a length-thickness ratio of 26 to 1 to take a 3° deflection without
breaking and the microscopic evidence does not indicate the presence of any
such long, thin pieces of quartz attached to the plate.

Although about a half-minute's etching with 48% hydrofluoric acid removes all quartz misoriented by more than 45', quartz misoriented by a smaller angle than this is not entirely removed by more than an hour's etching, as indicated by photographs taken with X-rays passing through the plate, a technique to be described later in this paper. The ionization chamber and amplifier are not sensitive enough to register the reflection from the small amount of this material left after two minutes' etching.

The disappearance of the more widely misoriented material in the earlier stages of etching may mean either that this material is preferentially removed or that there is uniform removal of all the misoriented material with the consequent disappearance of that which is smallest in amount. Geiger-Müller counter measurements of the intensity-distribution of the reflections from the misaligned material at various angles at the various stages of etching would indicate which of the two alternatives is true. These measurements are being made by Davisson and Haworth, but are not yet complete.

2.4 Photography of the Disturbed Surface with a Broad X-Ray Beam

A second photographic method with the single-crystal spectrometer, used by Berg in 1931[9], Gogoberidze in 1940[10], and others involves the reflection of a broad monochromatic beam from an appreciable area of the crystal surface (placed at the Bragg Angle, θ) onto a photographic plate or film placed parallel to the crystal face. The different reflection-intensities from the variously disturbed parts of the surface of the crystal plate darken the film differently, giving a map-like picture of the distribution of different degrees of disturbance over the surface of the plate. One picture produced in this way by Berg is reproduced in Fig. 7. The thin white cross and circle are reference marks scratched on the surface of the rock-salt crystal, the lines being parallel to the cube edges. The two sets of sub-parallel streaks are the traces of dodecahedral {101} planes and "show that the crystal structure in these places differed from the ideal lattice". They are interpreted as slip planes (störebenen) in the crystal. C. S. Barrett[11] has recently refined this technique and broadened its application to the study of a wide variety of metallurgical problems.

The application of this technique to the study of quartz surfaces might provide useful information on disturbance distribution which is not furnished by the other techniques.

[9] Berg, Wolfgang, "Über eine röntenographische Methode zur Untersuchung von Gitterstörungen an Kristallen," *Naturwissenschaften* 19 (1931), pp. 391–396.

[10] Gogoberidze, D. B., "Investigation of Surface Structure of Crystals by Means of Reflection of a Monochromatic X-Ray Beam, *Jour. Exptl. Physics, U.S.S.R.* 19 (1940) p. 96 (in Russian).

[11] C. S. Barrett; "A New Microscopy and Its Potentialities," *Metals Technology*, April, 1945.

3.1 The Double Crystal Spectrometer

In the double crystal spectrometer (Fig. 8), X-rays reflected from a crystal plate are again reflected from a second crystal plate into the ionization chamber.   Their intensity is indicated by a meter showing the amplified ionization current, as in the case of the single crystal spectrometer.   The divergent white rays from the collimating slits are reflected from crystal plate A as shown in Fig. 8: those of longer wave-length at higher angles and those of shorter wave-length at lower angles in accordance with Bragg's law.   If

S_1 S_2 $n\lambda = 2d \sin \theta$

θ_1

CRYSTAL
PLATE

θ_2

A

θ_3

λ_1 λ_3

ACTION OF CRYSTAL A λ_2

S_1 S_2

A

θ_1 θ_3 λ_1

θ_1

θ_3

B λ_3

ACTION OF A AND B TOGETHER
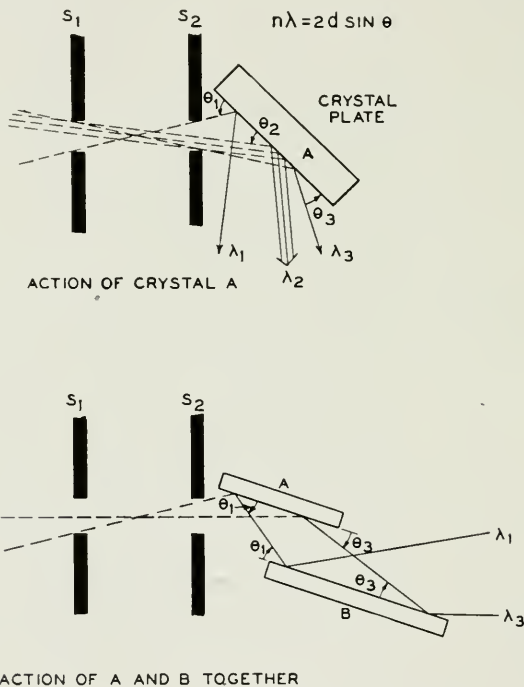
Fig. 8.—Double crystal spectrometer, parallel position. S_1 and S_2 are collimating slits

crystal plates A and B are placed so that similar sets of atomic planes in the two plates are parallel, the same rays that were reflected from plate A will also reflect from plate B as shown in Fig. 9 since each ray will meet this plate at the particular angle, θ, which will satisfy Bragg's law for that wave-length for the atomic planes in question.

If plate B is rotated a few seconds away from this position, however, and if the crystal is perfect, the conditions for reflection are destroyed for all rays so that no X-rays enter the ionization chamber.   However, a plate with a surface layer of misoriented crystal material will still reflect when thus

rotated because the misoriented particles will be brought into the reflecting position as the main plate is turned away from it. The farther the main plate is turned from the reflecting position, the weaker will be the reflected radiation because less quartz will be misoriented to this angle. In this respect the double crystal spectrometer technique is similar to C. J. Davisson's photographic technique, but is measuring much smaller angular rotation and higher reflection-intensity. A curve of the variation of reflection-intensity with angular rotation of the B crystal plate for two differently finished crystal plates measured by Davis and Stempel[12] is shown in Fig. 9
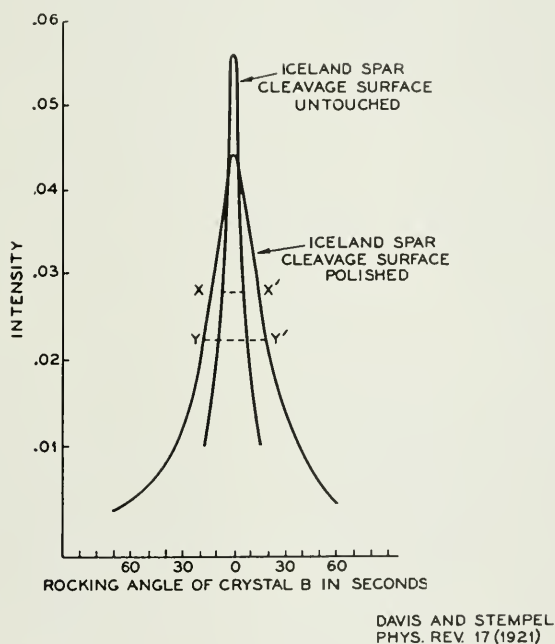
Fig. 9.—Double crystal spectrometer rocking curves

The higher reflection-intensity at the "zero angle" and the very rapid decrease of intensity as the plate is turned away from this position show that the untouched surface is less disturbed. The lower reflection-intensity at the "zero angle" and the less rapid decrease of intensity as the plate is turned away from this position show the polished surface to be more disturbed. The width of such "rocking" curves at half-maximum (as x-x' and Y-Y' in Fig. 9) has often been taken as a measure of perfection of the reflec-

[12] Davis, B. and Stempel, W. M., "An Experimental Study of the Reflection of X-Rays from Calcite," *Phys. Rev.* 17 (1921) pp. 608–623.

ting crystal. Bozorth and Haworth[13] made such measurements for variously prepared surfaces of quartz and found that the rocking-curve width at half-maximum was least for etched plates, the smallest width measured being 1.7 seconds. The greatest rocking-curve width measured was 88 seconds for which both crystal plates (A and B, Fig. 8) were lapped with 100-mesh carborundum. Bozorth and Haworth also showed that the rocking-curve width decreased very rapidly in the initial stages of etching

With this technique we are measuring the angular frequency distribution of the disturbed material that was detected with the intensity-ratio measure-measurements with the single crystal spectrometer. Thus, with the rocking curve, we show the intensity for each angle of incidence separately, but with the single crystal spectrometer we measure the intensity for a relatively wide range of angles of incidence at one measurement, which is the integrated intensity under the double crystal spectrometer rocking curve.

None of this small-angle disturbance is detectable by the Davisson photographic technique because at small angles the reflected rays are too close to those from the main plate and become confused with them on the film. Conversely, none of the material detected by the Davisson photographic technique is detectable by this technique because the intensity of the reflected rays from large-angle disturbance is not great enough to produce a measurable current in the ionization chamber.

Together, these various measurements indicate a relatively large amount of quartz misoriented by less than a minute and a smaller amount of quartz misoriented by larger angles up to three or four degrees.

The use of a photographic film in place of the ionization chamber in the double crystal spectrometer would presumably permit the precise measurement of the smaller amount of material misoriented by more than that now measured with this instrument, as in the Davisson technique with the single crystal spectrometer. This has not, to the writer's knowledge, been done.

4.1 THE TRANSMISSION LAUE CAMERA

In the transmission Laue camera (see Fig. 10) a beam comprising a large range of wave-lengths is passed through a stationary plate. Various sets of atomic planes in the crystal, each with a different interplanar spacing, d, making a variety of angles, θ, with the incident beam, reflect whatever wave-lengths of radiation in the incident beam satisfy the Bragg equation, $n\lambda = 2d \sin \theta$, and the reflected beams are recorded photographically. Figure 11 shows films resulting from one-hour exposures. As in the case of the single-crystal spectrometer, the slit-collimated beam comprises non-parallel rays,

[13] Bozorth, R. M. and Haworth, F. E., "The Perfection of Quartz and Other Crystals and Its Relation to Surface Treatment," *Phys. Rev.* 45 (1934) p. 821–826; *Bell Telephone System Technical Publications Monograph* B-801.
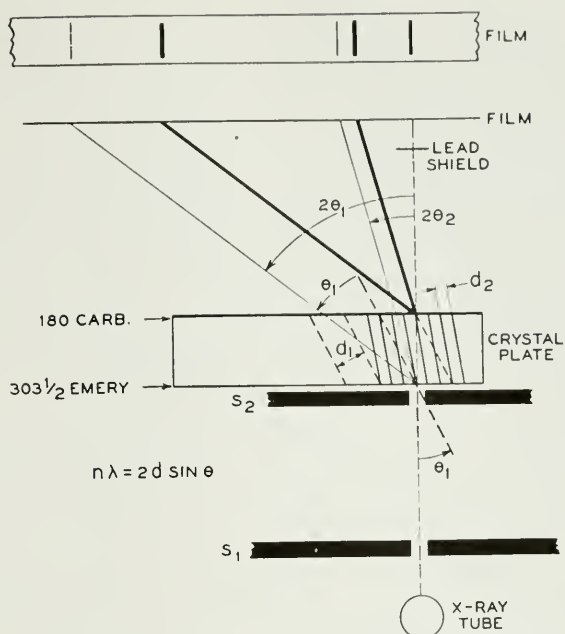
Fig. 10.—Section of transmission Laue camera. S_1 and S_2 are collimating slits
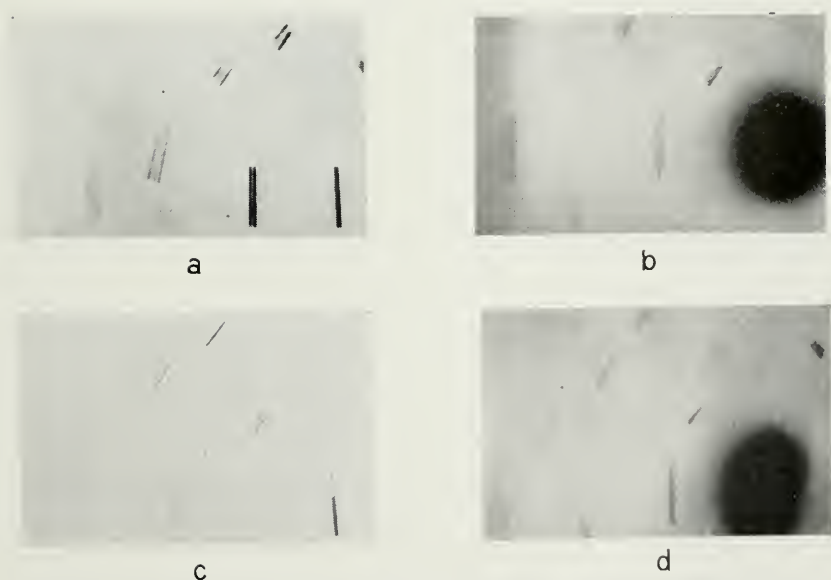
Fig. 11.—Transmission Laue photographs of a 4 mm.-thick BT quartz plate. (a) Face nearer film lapped with 180 carborundum. Other face lapped with $303\frac{1}{2}$ emery. (b) Same as a, after 47 hours' etching. (c) Same as a, but through a different part of the plate. Plate rotated 180° in its own plane. (d) Same as c, after 20 minutes' etching.

but in this case exposures are long enough to record the "white" radiation reflected from both the small-angle disturbed and undisturbed material. If reflections from material misoriented by more than a few minutes were recorded on the film the lines from the reflected X-rays would be broader than they are.

In the undisturbed material there is one wave length from a ray traveling in a given direction that will satisfy the Bragg equation for a given set of atomic planes. Most of these "usable" X-rays are reflected by the first thin layer of undisturbed crystal they meet and therefore very little reflection of X-rays of this wave-length from this ray takes place from deeper layers of the undisturbed crystalline material. This removal of the reflectable X-rays by the first thin layer of undisturbed crystal is known as "primary extinction".

In the disturbed surface layer, on the other hand, regions of dissimilar orientation are superposed. In this case a ray traveling in a given direction will have X-rays of one wave-length subtracted from it by the first crystalline material of a particular orientation it meets in accordance with the Bragg equation; then, beneath this, crystalline material at a different orientation will subtract from it X-rays of a different wave-length and so on so that from each ray a broader range of wave-lengths is diffracted by the disturbed material than by the undisturbed, resulting in a stronger reflected beam from the disturbed material. Since the lapped surfaces of the crystal plate give a stronger reflected beam than the undisturbed interior of the plate, the reflection of the slit collimated beam from each set of atomic planes appears on the film as a pair of lines. The density of these lines is related to the disturbance and their width is related to the depth of the disturbed surface layer, as shown diagrammatically in Fig. 10. The four photographs in Fig. 11 were taken in this way, all through the same BT quartz plate. Figure 11a shows the reflections from the various atomic planes of the 4 mm.-thick BT-cut quartz plate, lapped on one side with coarse abrasive (180 carborundum) and on the other with fine abrasive ($\#303\frac{1}{2}$ emery). As shown in Fig. 10, the coarsely lapped surface was toward the photographic film and therefore the line closer to the line from the direct beam (the single line in the lower right corner) is the stronger of the two. Figure 11b was taken in the same way after the plate had been etched in 48% hydrofluoric acid for 47 hours. The acid was renewed every few hours. The presence of disturbed material near the two surfaces is still discernible. Micrometer measurements after etching indicated that the thickness of the plate had been reduced by 0.14 mm.

Such a measurement is from the peaks of the rugged etched surface on one side of the plate to the peaks on the other side: over most of the plate the etching had proceeded to a greater depth than the .07 mm. indicated by the

micrometer measurements. Since the disturbance is still discernible it appears that its depth in this plate was greater than .07 mm.

With the discovery that the disturbed material may be more than 70 microns thick, it becomes apparent that great care must be taken to remove the disturbance produced by all previous lapping and sawing prior to measurement of disturbance produced by a particular lapping technique. This may require more than 48 hours' etching with 48% hydrofluoric acid which will produce a rugged surface. An alternative procedure is to use a natural crystal which has never been cut or worked. Since the natural faces of some crystals do show disturbance, preliminary tests should be made with the various techniques for detecting the presence of any disturbed material. A transmission Laue photograph of a small quartz crystal from Herkimer County, N. Y., taken by C. J. Davisson, showed no disturbed material. If a natural face of such a crystal were lapped under carefully controlled conditions and the resulting disturbance measured by the various techniques, a reliable picture of the disturbance produced by that lapping procedure would be obtained.

In many of these photographs some darkening occurs between the two lines that mark the outer surfaces of the plate. In most cases it appears as well-defined streaks, as in 11c, but may be irregular, as in 11a. Such streaks appear to be due to disturbed zones within the quartz plate whose disturbance is related to the growth history of the crystal. Where they adjoin the disturbed surface layer they may be responsible for erroneous measurements of the misorientation and depth of the surface layer.

Photograph 11c was taken prior to any etching, like 11a, but through a different part of the plate and shows different internal imperfections. Photograph 11d was taken through the same part of the plate after only 20 minutes' etching and shows very little surface disturbance. To get a picture of the distribution of the surface disturbance and internal imperfections of a plate a series of exposures should be taken with the plate translated a short distance relative to the beam between each exposure.

Measurements of the depth of the disturbed layer have given widely different results with different techniques. The Laue photographs just described indicate the depth to be more than .07 mm. or 70 microns in some cases, even with a $303\frac{1}{2}$ emery finish.[14] This is in accord with the 0.1 mm. depth assigned by Y. Sakisaka on the basis of two sets of measurements made by him with two widely different techniques.[15] This value is also in

[14] Since adequate precautions for the removal of all previous disturbance were not taken, this value may be erroneous.

[15] Sakisaka, Y., "The Effects of the Surface Conditions on the Intensity of Reflexion of X-rays by Quartz, *Japanese Journal of Physics* 4 (1927) p. 171–181.

Sakisaka, Y., "Reflexion of Monochromatic X-rays from Some Crystals," *Proc. Phys.-Math. Soc. of Japan*, Ser. III, v. 12 (1930), p. 189–202.

accord with the measurements on the double crystal spectrometer made by
Bozorth and Haworth who show that after 20 hours' etching with 48% hy-
drofluoric acid at 30° C. the rocking-curve width of a plate originally lapped
with 100-mesh carborundum was still measurably greater than that of a
plate originally lapped with 600-mesh carborundum.[13]

The measurements that have given half a micron for the depth of the dis-
turbed layer have been made with techniques incapable of showing part of
the disturbed material.   The reflection-intensity measurements made with
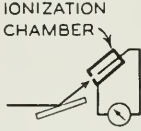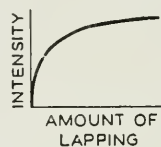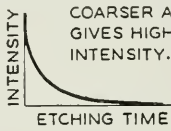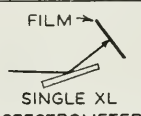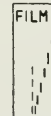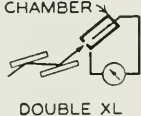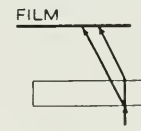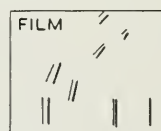
INSTRUMENT	TECHNIQUE	INFORMATION OBTAINED
IONIZATION CHAMBER \ SINGLE XL SPECTROMETER	REFLECTION– INTENSITY MEASUREMENT	COARSER ABRASIVE GIVES HIGHER INTENSITY.
FILM → SINGLE XL SPECTROMETER	PHOTOGRAPHY OF MISORIENTED MATERIAL	MATERIAL MISORIENTED UP TO 3°–6°, DEPENDING ON ABRASIVE
IONIZATION CHAMBER \ DOUBLE XL SPECTROMETER	ROCKING–CURVE MEASUREMENTS	COARSER ABRASIVE GIVES LOWER AND BROADER ROCKING CURVE WITH GREATER TOTAL INTENSITY. MAXIMUM QUARTZ MISORIENTATION MEASURED: ABOUT 1 MINUTE.
FILM LAUE CAMERA	PHOTOGRAPHY OF THICK PLATES	MOST REFLECTION FROM LAPPED SURFACES. COARSER ABRASIVE GIVES DARKER SPOT.

Fig. 12.—Diagrammatic summary of instruments, techniques, and results

the single crystal spectrometer can only show the material present in large
enough amount to cause measurable ionization in the ionization chamber.
The Davisson photographs with the single crystal spectrometer cannot show
material misaligned by less than 15 minutes.   Any photographic technique
which is capable of measuring material misoriented by a few seconds has
shown a disturbed layer much thicker than half a micron at any worked
surface of quartz crystal.

Figure 12 is a diagrammatic summary of the various techniques that have
been described.   Table 1 summarizes the present knowledge concerning the

TABLE I

Summary of Information Concerning the Nature of the Disturbed Surface
Material of Quartz Plates

Description of the disturbed material	Method of detection
Randomly oriented powder on the surface, removable by scrubbing	Electron diffraction photography only
Material misoriented from approximately 45′ to approximately $4\frac{1}{2}°$, removable by about 30 seconds' etching with 48% hydrofluoric acid. Not removable by scrubbing. About $\frac{1}{2}$ micron thick.	Photography with the single crystal spectrometer
Material misoriented from 0° to 45′, in some cases requiring more than 47 hours' etching with 48% hydrofluoric acid for removal. Not removable by scrubbing. May be 50–100 microns thick.	Reflection intensity measurements with the single crystal spectrometer. Rocking-curve width measurements with the double-crystal spectrometer. Laue photography of thick crystal plates.
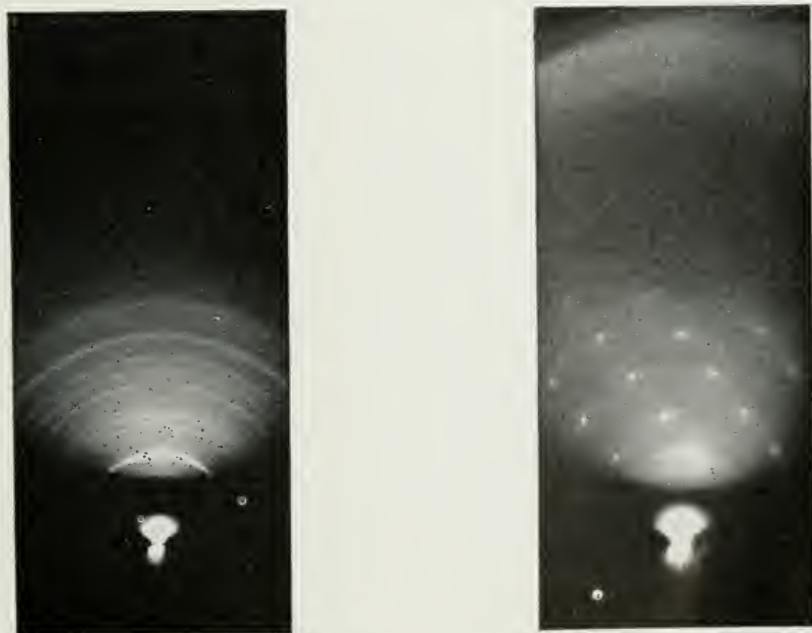
Fig. 13.—Electron diffraction photographs of a BT quartz plate taken with a 50 kv. electron beam (C. J. Davisson)

disturbed layer of worked surfaces of quartz, indicating the technique through which the information was obtained in each case. In order that this table shall be complete the electron diffraction work of C. J. Davisson

must be included although it does not properly belong in a paper on X-ray
techniques.   Dr. Davisson took an electron diffraction photograph of a
quartz plate that had been lapped with $303\frac{1}{2}$ emery, water rinsed, and air
dried.   The plate was then scrubbed vigorously with soap and water and
toothbrush and a second photograph was taken.   The two photographs are
reproduced in Fig. 13.   The first shows a series of continuous rings indicat-
ing the presence of a large number of small particles of quartz with random
orientation.   In the second photograph, these rings have disappeared and
there remain only arc-segments associated with spots.   The spots are the
"reflections" from undisturbed quartz.   The arcs represent quartz rotated
through a small range of angles from this position, the misoriented material
indicated by the Davisson photographs with the single crystal spectrometer.
These electron diffraction photographs show that a lapped plate has ran-
domly oriented quartz on its surface which may be removed by scrubbing
and quartz with limited misorientation which is not removed by scrubbing.
No X-ray technique has shown the existence of the randomly oriented
material.

BIBLIOGRAPHY

Allison, S. K.; "The Reflecting and Resolving Power of Calcite for X-Rays," *Phys. Rev.* 41
    (1932) pp. 1–20.
Barraud, J.; Structure complexe des taches dans les diagrammes de Laue, *Compt. rend.*
    217 (1943) pp. 394–396.
Barrett, C. S.; A New Microscopy and Its Potentialities, *Metals Technology*, April, 1945.
Barret, C. S.; "Laue Spots from Perfect, Imperfect and Oscillating Crystals," *Phys. Rev.*
    38 (1931) pp. 832-833.
Barret, C. S. and Howe, C. E.; "X-Ray Reflections from Inhomogeneously Strained
    Quartz," *Phys. Rev.* 39 (1932) pp. 889–897.
Berg, Wolfgang; "Über eine Röntgenographische Methode zur Untersuchung von Gitter-
    störungen an Kristallen," *Naturwissenschaften* 19 (1931) pp. 391–396.
Bollman, V. L. and Du Mond, J. W. M.; "Evidence for Surface Layers in Cleaved Calcite
    Crystals," *Phys. Rev.* 54 (1938) pp. 238–239.   (Abstract)
Bozorth, R. M. and Haworth, F. E.; "The Perfection of Quartz and Other Crystals and
    Its Relation to Surface Treatment," *Phys. Rev.* 45 (1934) pp. 821–826.
Bragg, W. L. and Darwin, C. G.; "The Intensity of Reflexion of X-Rays by Crystals," *Phil.
    Mag.* 1 (Ser. 7) (1926) pp. 897–922.
Bragg, W. L., James, R. W. and Bosanquet, C. H.; "The Intensity of Reflexion of X-Rays
    by Rock Salt,"   Part I, *Phil. Mag.* 41 (1921) pp. 309–337.
Bragg, W. L., James, R. W. and Bosanquet, C. H.; "The Intensity of Reflexion of X-Rays
    by Rock Salt,"   Part II, *Phil. Mag.* 42 (1921) pp. 1–17.
Bragg, W. L. and West, J.; "A Technique for the X-Ray Examination of Crystal Structures
    with Many Parameters," *Zschr. f. Krist.* 69 (1929) pp. 118–149.
de Broglié, M.; "Sur la Réflexion des Rayons de Röntgen," *Comptes Rendus* 156 (1913) pp.
    1153–1155.
de Broglié, M. and Lindemann, F.-A.; "Sur les Phénomènes Optiques Présentés par les
    Rayons de Röntgen rencontrant des Milieux Cristallins," *Comptes Rendus* 156 (1913)
    pp. 1461–1463.
Colby, M. Y. and Harris, S.; "Effect of Etching cn the Relative Intensities of the Compo-
    nents of Double Laue Spots Obtained from a Quartz Crystal," *Phys. Rev.* 43 (1933)
    pp. 562–563.
Compton, A. H.; "The Reflection Coefficient of Monochromatic X-Rays from Rock Salt
    and Calcite," *Phys. Rev.* 10 (1917) pp. 95–96.
Darwin, C. G.; "The Reflexion of X-Rays from Imperfect Crystals," *Phil. Mag.* 43 (1922)
    pp. 800–829.

Davis, Bergen and Stempel, W. M.; "An Experimental Study of the Reflection of X-Rays from Calcite," *Phys. Rev.* 17 (1921) pp. 608–623.

Dickinson, R. G. and Goodhue, E. A.; "The Crystal Structure of Sodium Chlorate and Sodium Bromate," *Jour. Amer. Chem. Soc.* 43 (1921) pp. 2045–2055.

Du Mond, V. L. and Bollman, J. W. M.; "New and Unexplained Effects in Laue X-Ray Reflections in Calcite," *Phys. Rev.* 50 (1936) p. 97.

Du Mond, V. L. and Bollman, J. W. M.; "Test of the Validity of X-Ray Crystal Methods of Determining *e*," *Phys. Rev.* 50 (1936) pp. 524–537.

Ehrenberg, W. and Mark, H.; "Über die natürliche Breite der Röntgenemissionslinien," *Ztschr. f. Physik* 42 (1927) pp. 807–822 and pp. 823–831.

Ewald, P. P.; "Die Intensitäten der Röntgenreflexe und der Strukturfaktor," *Phys. Ztschr.* 26 (1925) pp. 29–32.

Finch, G. I.; "The Beilby Layer on Non-Metals," *Nature* 138 (1936) p. 1010.

Finch, G. I.; "The Nature of Polish," *Trans. Faraday Soc.* 33 (1937) pp. 425–430.

Gogoberidze, D. B.; "Investigation of Surface Structure of Crystals by Means of Reflection of a Monochromatic X-Ray Beam," *Jour. Exptl. Phys. U.S.S.R.* 10 (1940) pp. 96–103 (in Russian).

Hirsch, F. R. and Du Mond, J. W. M., "X-Ray Evidence on the Nature of the Surface Layers of Thin Ground Quartz Crystals Secured with the Cauchois Spectrograph," *Phys. Rev.* 54 (1938) pp. 789–792.

Hirsch, F. R.; "The Beilby Layer on Thin Ground Quartz Crystals," *Phys. Rev.* 54 (1938) p. 238.

Hirsch, F. R.; "Note on Thin Quartz Crystals as used in the Cauchois Focusing X-Ray Spectrograph," *Phys. Rev.* 58 (1940) pp. 78–81.

Joffé, A. F.; "The Physics of Crystals," McGraw, New York, 1938, p. 38.

Kirkpatrick, P. and Ross, P. A.; "Absolute X-Ray Reflectivities of Single Crystals of Calcite, Rocksalt, Rochelle Salt, and Barite," *Phys. Rev.* 43 , pp. 596–600.

Manning, K. V.; "The Effect of Treatment of the Surfaces of Calcite Crystals upon the Resolving Power of the Two-Crystal Spectrometer," *Rev. Sci. Instr.* 5 (1934) pp. 316–320.

Murdock, C. C.; "Multiple Laue Spots," *Phys. Rev.* 45 (1934) pp. 117–118.

Richtmeyer, F. K., Barnes, S. W. and Manning, K. V.; "The Effect of Grinding and Etching on a Pair of Split Calcite Crystals," *Phys. Rev.* 44 (1933) pp. 311–312.

Sakisaka, Y.; "The Effects of the Surface Conditions on the Intensity of Reflexion of X-Rays by Quartz," *Japanese Jour. Phys.* 4 (1927) pp. 171–181.

Sakisaka, Y.; "Reflexion of Monochromatic X-Rays from Some Crystals," *Proc. Phys.-Math. Soc. of Japan,* Series III, v. 12 (1930) pp. 189–202.

Sakisaka, Y. and Sumoto, I.; "The Effects of the Thermal Strain on the Intensity of Reflexion of X-Rays by Some Crystals," *Proc. Phys.-Math. Soc. of Japan,* Series III, v. 13 (1931) pp. 211–217.

Wagner, E. and Kulenkampf, H.; "Die Intensität der Reflexion von Röntgenstrahlen verschiedener Wellenlänge an Kalkspat und Steinsalz," *Annalen der Physik* 68 (1922) pp. 369–413.

NOTE

Some Novel Expressions for the Propagation Constant
of a Uniform Line

By J. L. CLARKE

IN a modern communication system the leakance is generally quite
small and can be neglected in making propagation computations.

Hence the case is taken of a circuit having resistance, inductance and
capacitance.

Let R, L and C be the resistance, inductance and capacitance per mile
of the circuit.

Now the formulae which have been developed for the real and imaginary
parts of P, the propagation constant, are:

$$P = \alpha + j\beta$$

where

$$\alpha = \sqrt{\tfrac{1}{2}\omega C\{\sqrt{R^2 + \omega^2 L^2} - \omega L\}} \qquad (1)$$

$$\beta = \sqrt{\tfrac{1}{2}\omega C\{\sqrt{R^2 + \omega^2 L^2} + \omega L\}} \qquad (2)$$

Inspection of these formulae does not reveal any obvious reason for the
particular structure of the expressions; however, they may be changed so
as to present themselves in a different form.

P can be written:

$$P = \beta(\alpha/\beta + j) \qquad (3)$$

Now

$$\frac{\alpha}{\beta} = \sqrt{\frac{\sqrt{R^2 + \omega^2 L^2} - \omega L}{\sqrt{R^2 + \omega^2 L^2} + \omega L}} = \tan\theta \qquad (4)$$

where θ is the absolute value of the angle of the characteristic impedance.

Hence

$$P = \beta(\tan\theta + j) \qquad (5)$$

The expression in (4) for tan θ may be written:

$$\tan \theta = \sqrt{1 - LCV^2} \qquad (6)$$

where V = the phase velocity.
Now $LC = 1/V_0^2$
where V_0 = the transient velocity.
Hence

$$\tan \theta = \sqrt{1 - V^2/V_0^2} \qquad (7)$$

and

$$P = \beta(\sqrt{1 - V^2/V_0^2} + j) \qquad (8)$$

Now it can also be shown that for a progressive wave:

$$\sqrt{1 - \frac{V^2}{V_0^2}} = \sqrt{\frac{\frac{1}{2}CE^2 - \frac{1}{2}LI^2}{\frac{1}{2}CE^2 + \frac{1}{2}LI^2}} \qquad (9)$$

So that the real part of the propagation constant is equal to the square root of the difference between the electrostatic energy per mile and the electromagnetic energy per mile divided by the sum of the energies (multiplied by β).

This gives the attenuation constant in terms of the most fundamental entity that we know, namely energy.

Also multiplying each side of equation (8) by x the geographical length of the line between the origin and the point under consideration:

$$Px = \beta(\sqrt{1 - V^2/V_0^2}\, x + jx) \qquad (10)$$

Now βx is the number of radians (i.e. total phase shift along the length x) and using relativity conceptions,

$$x \sqrt{1 - \frac{V^2}{V_0^2}}$$

is the apparent length of the wave structure in the length x (the waves moving at velocity V) as observed from a fixed point with signals moving at velocity V_0.

Also $\beta x \sqrt{1 - V^2/V_0^2}$ is the number of radians in the apparent length $x\sqrt{1 - V^2/V_0^2}$

The characteristic impedance may be stated in terms of the phase velocity as follows:

$$Z_0 = \frac{1}{CV} - j\,\frac{RV}{2\omega} \qquad (11)$$

NOTE

Decibel Tables

By K. S. JOHNSON

ABOUT twenty years ago, Bell System communication engineers felt the need for, and adopted, a new "Transmission Unit," initially abbreviated "TU" but some five years later given the name "Decibel," or db. This unit is now universally used throughout the communication world and is fundamentally a measure of the ratio of any two powers. Quantitatively, the number of decibels corresponding to the ratio of any two powers is 10 times the common logarithm of that ratio.* If the ratio of the first power to the second power is greater than unity, the first power is said to represent a transmission "gain" with respect to the second power, or the latter is said to represent a "loss" with respect to the first power.

Since currents flowing through, or voltages impressed across, the same or equal impedances will result in powers that are proportional to the squares of these currents or voltages, it is possible, under these specific conditions, to state that the number of decibels corresponding to the ratio of any two such currents or voltages is 20 times the common logarithm of the absolute magnitude of the ratio of these currents or voltages.

Another unit that is frequently employed in theoretical transmission problems is the "Neper." The use of this unit often results in the simplification of such problems and, hence, its relationship to the decibel and to exponential and hyperbolic functions is frequently of interest to communication engineers.

Although the relations between these various values are obviously not complicated ones, it has been found by experience that tables of numerical values are often very useful to communication engineers. As a result, rather extended tables (21 pages) have been computed under the direction of the writer and P. H. Richardson, in which the entering arguments are: (1) decibels, with the tabular values giving the corresponding current, voltage, or power ratios—and their reciprocals; (2) current or voltage ratios, with the tabular values being the corresponding decibels. Tables (16 pages) have also been computed in which the entering arguments are decibels and the corresponding tabular values are nepers (A), e^A, e^{-A}, sinh A, cosh A and tanh A. These latter tables are, among other things, useful in the design of attenuators or pads, etc.

Photo offset copies of any of the above tables may be obtained gratis from the Director of Publication of the Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y.

---

* See "Decibel—The Name for the Transmission Unit", by W. H. Martin, *Bell Sys. Tech. Jour.*, January 1929.

Abstracts of Technical Articles by Bell System Authors

Network Analysis and Feedback Amplifier Design.[1] H. W. BODE. The material for this book was originally prepared as a text for an informal course at Bell Telephone Laboratories. It is the outgrowth of a research directed at the problem of designing degenerative feedback amplifiers to provide substantial feedback without instability. The solution of the feedback problem is, however, dependent upon certain propositions in general network theory which are applicable also to other situations. With the addition of other logically related material, this has made the book primarily a text on general network theory.

Earlier texts on networks have been concerned primarily with transmission line and filter theory. The present book emphasizes the broadband aspects of network theory. In other words, it is concerned with the problem of providing characteristics which vary smoothly, and in some prescribed manner over a broad frequency range. This aspect of network theory is stressed because it is the one which best fits the feedback problem. It also has applications, however, to the many broad-band problems which arise in television, frequency modulation, multi-channel carrier telephone and other modern communication systems.

The emphasis on broad-band problems has a number of consequences. For example, it gives special importance to networks including resistances as well as reactances, since it is frequently only by the use of controlled dissipation that network characteristics can be made to vary smoothly over broad ranges. The emphasis on broad-band applications also requires special attention to the effects of parasitic elements, and several sections of the book are devoted to the development of design methods for networks including prescribed parasites. A final consequence is the importance which is assumed by the limitations on the characteristics which can be obtained from physical networks. Over very narrow bands only very mild limitations exist, but as the band becomes broader the available characteristics become more and more restricted.

The other principal point of emphasis of the book is on the use of networks in association with vacuum tubes, rather than as purely passive structures. The primary theoretical development of the book is stated in terms of general active circuits. Otherwise the effort to extend network

[1] Published by D. Van Nostrand Company, Inc., New York, N. Y., 1945.

159

theory to vacuum tube circuits consists chiefly in giving special emphasis to network design problems ordinarily found as part of vacuum tube amplifier design. The design of an over-all feedback loop is, of course, an outstanding example. In addition, special attention is also given to the design of such individual network units as input and output circuits, interstage networks, and local feedback circuits, especially when they appear as constituents of a broad-band amplifier.

Judging Mica Quality Electrically.[2] K. G. COUTLEE. A threatened mica shortage resulting from an unprecedented wartime demand for mica capacitors used in electronic communication equipment by the Armed Forces was forestalled by rigid conservation measures, use of alternate materials, and the use of electrically selected mica from types previously considered unsuitable for capacitor use. By employing two electrical tests, developed by Bell Telephone Laboratories, Inc. for the War Production Board, in combination with visual and physical requirements, mica was selected from plentiful stocks of lower visual quality types of mica, effectively increasing the supply of capacitor mica by 60 per cent. This method of electrically judging the quality of raw mica was given a thorough commercial trial and found both practicable and reliable.

A Simple Optical Method for the Synthesis and Evaluation of Television Images.[3] R. E. GRAHAM and F. W. REYNOLDS. A combination of a 35-millimeter motion-picture projector and a line screen enables the projection of still or motion pictures closely similar in appearance to those produced by television. This similarity of appearance is checked theoreticallly by an analysis of the type previously reported by Mertz and Gray in a discussion of the theory of scanning. From the analysis it is shown that five parameters of the optical-simulation system may be varied to obtain the equivalent of variations in television factors such as number of scanning lines, size and configuration of scanning apertures, and width of frequency band.

Photographs of simulated television pictures projected by this method are presented. These pictures include subject matter of general interest as well as as selected subjects to illustrate the spurious detail components introduced by the television scanning process. These components produce moiré patterns, "steps" on diagonal lines, and impairment of vertical resolution. Simulation pictures projected by this method have been compared with those produced by a television system and the expected agreement observed.

Calculations are given of the diffraction effects in optical systems of this type and it is shown that the departure from geometrical theory is small in the arrangements described.

[2] *Elec. Engg., Trans. Sec.,* November 1945.
[3] *Proc. I. R. E.* and *Waves and Electrons,* January 1946 (pp. 18W–30W).

A Coil-Neutralized Vacuum-Tube Amplifier at Very High Frequencies.[4] R. J. KIRCHER. This paper describes a two-stage single-side coil-neutralized amplifier employing an experimental triode operating in the vicinity of 140 megacycles. Circuit features are described and typical operating conditions are indicated. Typical distortion characteristics at low-power levels are also included.

Fundamental Theory of Servomechanisms.[5] LEROY A. MACCOLL. The use of servomechanisms and related devices for automatic control and regulation is very old, dating back to the latter part of the eighteenth century. However, it is only recently, approximately since the beginning of the war, that it has been recognized that these devices are essentially feedback amplifiers in a mechanical, or partly mechanical, form. From the recognition of this fact it follows that the highly developed theory of electrical feedback amplifiers can be applied at once to servomechanisms and similar devices.

This book, which was originally intended to be a National Defense Research Committee report, is an introduction to the theory of linear servomechanisms, considered as a special application of the general theory of feedback amplifiers. The steady-state theory of the systems is taken as fundamental, and the various problems concerning the stability and performance of the systems are discussed in terms of it. In the several chapters a variety of types of linear servomechanisms are considered. A brief discussion of one simple non-linear servomechanism is given in the Appendix.

Corrosion Protection for Transcontinental Cable West of Salt Lake City, Utah.[6] T. J. MAITLAND. This paper discusses the problems involved in maintaining the effectiveness of the thermoplastic covering provided on buried toll cables for installation in areas where corrosion is anticipated. It also describes the method used to obtain the required supplemental electrical drainage for the Transcontinental Cables across the Great Salt Desert west of Salt Lake City where the low earth resistivity and high concentration of alkali salts preclude the use of rectifiers connected between cable sheath and a made ground generally employed for drainage purposes. Such installations would result in negative potentials between cable and earth of sufficient magnitude to create conditions conducive to cathodic corrosion of the lead sheath in the presence of an alkali salt electrolyte.

To provide electrical drainage without incurring these excessive negative potentials a method was developed utilizing the normal potential difference between zinc and lead as the source of drainage current. Twenty-four pound zinc bars of commercially available zinc, 99 per cent pure, were installed directly in the ground a short distance from the cables at 12-mile

[4] *Proc. I. R. E.*, December, 1945.
[5] Published by D. Van Nostrand Company, Inc., New York, N. Y., 1945.
[6] *Corrosion*, June 1945.

intervals, making connection between the zinc anodes and the cable sheaths by buried wire. The cable-to-earth potentials were appreciably affected throughout the entire 120 route miles across the Great Salt Desert by this procedure.

During the year these anodes have been in place, the cables have remained at a satisfactory negative potential to earth (.20 to .50 volt) with a small current being constantly drained to the zinc anodes. It is considered from the results to date that for similar areas the use of metallic anodes offers an economical and satisfactory means for protecting buried cables against corrosion.

Transmission Networks for Frequency Modulation and Television.[7] HAR-OLD S. OSBORNE. Looking forward to a great post war expansion in the arts of frequency modulation and television this paper discusses plans of the Bell System for providing transmission networks required for the interconnection of broadcast stations. A review of cable and open-wire carrier systems shows how developments for purely message telephone business have at the same time put the Bell System in a position of being able at the present time to meet such network transmission requirements for frequency modulation as the broadcasters may select as desirable. Coaxial developments are reviewed briefly, including the application of these developments to television transmission. Future developments, together with the coaxial construction plans now under way, are expected to provide by about 1950 a fairly comprehensive nationwide network of facilities capable of providing for such transmission requirements as may be desired by the television industry. The important features involved in the operation of such networks are discussed, indicating a requirement for a highly trained nationwide organization and much equipment—a requirement which the Telephone Companies can face with confidence because of their experience in handling nationwide communications.

Visible Patterns of Sound.[8] RALPH K. POTTER. New ways of translating sounds into pictures are described. These methods of sound portrayal are unique because what may be seen in the sound patterns is consistent with what is heard in the original sound. The pictures display the three basic dimensions of sound-pitch, loudness and time—in a form somewhat analogous to a musical score. Experimental training has shown that with practice one may learn to read such patterns of speech so that the development offers the ultimate possibility of aid to the severely deafened in learning to speak correctly and to use the telephone by seeing rather than hearing the voice of the distant speaker. The patterns will also be of considerable interest in the fields of speech science and music.

[7] *Elec. Engg.*, November 1945.
[8] *Science*, November 9, 1945; *Bell Tel. Sys. Monograph* B-1368.

General Formulas for "T"- and "Π"- Network Equivalents.[9] MYRIL B. REED. This paper presents the development of two sets of general formulas which determine a set of "T" or "Π" impedances equivalent to any linear, lumped-constant, four-terminal network.

Concerning Hallén's Integral Equation for Cylindrical Antennas.[10] S. A. SCHELKUNOFF. The main purpose of this paper is to explain the substantial quantitative discrepancy between Hallén's formula for the impedance of cylindrical antennas, and ours. Hallén's first approximation involves a tacit assumption that the antenna is short compared with the wavelength. Since the subsequent approximations depend on the first, they are degraded by this initial assumption.

The approximations involved in his integral equation itself are justified; and, if properly handled, the equation yields results in much better agreement with ours. The last section of the paper is devoted to infinitely long antennas. Such antennas can be treated by at least three very different methods and a comparison is instructive. In practice, the solution for this case is an approximation to a long antenna designed to carry progressive waves.

Principal and Complementary Waves in Antennas.[11] S. A. SCHELKUNOFF. In response to an increased interest in mathematical aspects of antenna theory, this paper presents details of analysis of cylindrical and other non-conical antennas as a supplement to a previous paper containing the outline of the method and the main results. In the course of the present discussion the theory of principal waves on cylindrical conductors is extended to include the case in which the diameter is not small compared with the wavelength.

Research Revolutionizes Materials.[12] J. R. TOWNSEND. A technological lesson to be drawn from defeated Germany is that whereas Germans had been noted for their fundamental contributions to science, they were unable to compete with the United Nations in the field of applied science and particularly in high-speed production methods. Their defeat was due more to the overwhelming number than to the individual superiority of the arms brought against then. The miracle of American production is based on a design related to obtaining the most from the process used, materials of uniform quality, and high-speed production methods using high-power automatic machinery. Germany's failure was due to standardizing too early and too inflexibly and this meant that they could not compete with the steady improvements in the art. The usual procedure is the development of methods of test followed by collection of data and the formulation of specific requirements controlling the useful quality of the material. Modern

[9] *Proc. I. R. E.*, December, 1945.
[10] *Proc. I. R. E.*, December, 1945.
[11] *Proc. I. R. E.*, January, 1946.
[12] *A.S.T.M. Bulletin*, December, 1945.

industry is based upon such specifications because materials must be so controlled since the action of the machine is unvarying. Modern statistical methods can be applied to provide the tolerances and allowances necessary to achieve a uniform product. The work of the American Society for Testing Materials broadly covers the field of research in materials, methods of test, and quality control. The benefits of this work extend to vast improvement in process methods, more uniform and higher-quality material and result in economic gains of extensive character. Three examples were cited illustrating extensive projects of great use to the war effort. These were the development of requirements for sheet brass, which was applied specifically to production of cartridge cases, high-quality die-casting specifications resulting in the production of many parts used in communication and aviation equipment, and the development of a method of test for inspecting mica by an electrical rather than a visual test. This last resulted in a large economic saving of this scarce material.

Infantry Combat Communications.[13] RALPH E. WILLEY. Communications within an infantry division during combat involve not only the efficient installation, operation and maintenance of all means of communication normally provided and adopted for specific functions but also the use of standard equipment in improvised methods adapted to the needs of the particular situation. The paper covers a brief description of the major items of signal equipment issued to an infantry division together with their normal use. In addition, there is discussed the solution to many field problems based on the combat experience of the writer in Belgium, Holland and Germany.

Interesting information is given on the signal supply problem and combat losses over a six-month period of combat. Improvised field radio-link installations and remote controls for the protection of operating personnel are discussed briefly. Photographs included with the paper show pictorially the majority of the items of equipment described.

[13] *Elec. Engg.*, January, 1946.

Contributors to this Issue

Elizabeth J. Armstrong, M.A. in Geology, Bryn Mawr College, 1934; Ph.D., 1939. Lecturer in Geology, Barnard College, 1938–41; Research Assistant, Columbia University 1941–42; National Research Council Fellow, 1942–43; Bell Telephone Laboratories, 1943–. Recently Miss Armstrong has been working on X-ray studies of crystal strain and imperfections. X-ray irradiation and twinning.

J. L. Clarke, B.Sc. in Electrical Engineering, University of London, 1909. Bell Telephone Company of Canada, 19_0–; Transmission Engineer, 1924–.

John W. Clark, A.B., University of Montana, 1935; M.S., University of of Illinois, 1937; Ph.D., 1939. Dr. Clark joined the Technical Staff of the Bell Telephone Laboratories in 1939, where his principal interest has been in the development of vacuum tubes for use at ultra-high frequencies.

William H. C. Higgins, Purdue University, B.S. 1929; E.E. 1934. Development and Research Department, American Telephone and Telegraph Company, 1929–34; Bell Telephone Laboratories, 1934–. During his employment with the A. T. & T. Company Mr. Higgins was engaged in carrier telephone and program transmission studies. Since his tranfer to Bell Telephone Laboratories he had been concerned with radio and radar development.

Kenneth S. Johnson, A.B., Harvard University, 1907; Harvard Graduate School of Applied Science, 1907–09. Transmission and Protection Department, American Telephone and Telegraph Company, 1909–13; Engineering Department, Western Electric Company, 1913–25; Bell Telephone Laboratories, 1925–. Formerly as Transmission Research Engineer and now as Transmission Standards Engineer, Mr. Johnson has long been engaged in work closely connected with all types of transmission network problems. He is the author of the book "Transmission Circuits for Telephonic Communication."

John Leutritz, B.S. in Chemistry, Bowdoin College, 1929; A. M. in Botany, Columbia University, 1934; Ph.D., Columbia University, 1945. U. S. Navy, Medical Corps, 1921–25. Bell Telephone Laboratories, 1929–. Mr. Leutritz' interest has been along biological lines, primarily in respect

to wood preservation.   During the war period he was active in the protection of electrical equipments against moisture and fungi attack when used in the tropics.

WILLIAM W. MUMFORD, B.A., Willamette University, 1930.   Bell Telephone Laboratories, 1930–.   Mr. Mumford has been engaged in work that is chiefly concerned with ultra-short-wave and microwave radio communication.

A. L. SAMUEL, A.B., College of Emporia (Kansas), 1923; S.B. and S.M. in Electrical Engineering, Massachusetts Institute of Technology, 1926. Additional graduate work at M. I. T. and at Columbia University.   Instructor in Electrical Engineering, M. I. T., 1926–28.   Mr. Samuel joined the Technical Staff of the Bell Telephone Laboratories in 1928, where he has been engaged in electronic research and development.   Since 1931, his principal interest has been in the development of vacuum tubes for use at ultra-high frequencies.

WILLIAM C. TINUS, B.S. in Electrical Engineering, Texas A. & M. College, 1928.   Bell Telephone Laboratories, 1928–.   For ten years Mr. Tinus was engaged in the development of radio communication apparatus, principally for mobile use.   In 1938 he organized and directed the first radar development work in the Laboratories.   As Radio Development Engineer he is now responsible for a number of radar and related electronic development projects for the Army and Navy.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

THE BELL SYSTEM TECHNICAL JOURNAL

━━━━━━━━━

━━━━━━━━━

━━━━━━━━━

Extra copies of the MAGNETRON AS A GENERATOR OF CENTI-
METER WAVES by Fisk, Hagstrum and Hartman, the sole article com-
prising this issue of the Bell System Technical Journal are available at
$1.00 per copy.

Because circumstances have delayed the issuance of this number of the
Bell System Technical Journal far beyond its designated date of appearance,
it seems desirable to record that issuance actually occurred on July 16, 1946.

CORRECTION FOR ISSUE OF JANUARY, 1946

Page 35: Footnote, 5 "The Magnetron as a Generator of Centimeter Waves," J. B. Fisk, H. G. Hagstrum, and P. L. Hartman, *B. S. T. J.*, January, 1946.

should read, 5 "The Magnetron as a Generator of Centimeter Waves," J. B. Fisk, H. G. Hagstrum, and P. L. Hartman, to appear in *B. S. T. J.*, April, 1946.

The Magnetron as a Generator of Centimeter Waves

By J. B. FISK, H. D. HAGSTRUM, and P. L. HARTMAN

INTRODUCTION

L ATE in the summer of 1940, a fire control radar operating at 700 mega-cycles per second was in an advanced state of development at the Bell Telephone Laboratories. The pulse power of this radar was generated by a pair of triodes operating near their upper limit of frequency. Even when driven to the point where tube life was short, the generator produced peak power in each pulse of only two kilowatts, a quantity usable but marginal. Although the triodes employed had not been designed for high voltage pulsed operation, they were the best available. This is an example of how develop-ment of radar in the centimeter wave region was circumscribed by the lack of a generator of adequate power and reasonable life expectancy. More-over, the prospects of improvement of the triode as a power generator at these wavelengths and extension of its use to shorter wavelengths were not bright. Solution of the problem by means of power amplification was remote. A new source of centimeter wave power was urgently needed.

For the British, who were at war, the problem was even more urgent. They had undertaken a vigorous search for a new type of generator of sufficiently high power and frequency to make airborne radar practicable in the defense against enemy night bombers. They found a solution in the multiresonator magnetron oscillator, admirably suited to pulsed generation of centimeter waves of high power.

In the fall of 1940, an early model of this magnetron operating at ten cen-timeters was brought to the United States for examination. The first Ameri-can test of its output power capabilities was made on October 6, 1940 in the Bell Telephone Laboratories' radio laboratory at Whippany, N. J. This test confirmed British information and demonstrated that a generator now existed which could supply several times the power that our triodes delivered and at a frequency four times as great. The most important restraint on the development of radar in the centimeter wavelength region had now been removed.

A number of pressing questions remained to be answered, however. Could the new magnetron be reproduced quickly and in quantity? Was its

operating life satisfactory? Could its efficiency and output power be substantially increased? Could one construct similar magnetrons at forty centimeters, at three centimeters, even at one centimeter? Could the magnetron oscillator be tuned conveniently? One by one, during the war years, all of these questions have been answered in the affirmative. In many instances, but not without detours and delays, results have been better than expected or hoped for.

The British magnetron was first reproduced in America at the Bell Telephone Laboratories for use in its radar developments and those at the Radiation Laboratory of the National Defense Research Committee which was then being formed at the Massachusetts Institute of Technology. Since that time, extensive research and development work has been carried on in our Laboratories, in other industrial laboratories, and in the laboratories of the National Defense Research Committee. Several manufacturers have produced the resultant designs. Magnetron research and development was also carried on in Great Britain by governmental and industrial laboratories. There has been continuous interchange of information among all these laboratories through visits and written reports. Magnetron and radar developments have been greatly accelerated by this interchange.

Multicavity magnetron oscillators are now available for use as pulsed and continuous wave generators at wavelengths from approximately 0.5 to 50 centimeters. The upper limit of peak power is now about 100 kilowatts at 1 centimeter, 3 megawatts at 10 centimeters. Operating voltages may be less than 1 kilovolt or more than 40 kilovolts. The magnetic fields essential to operation range from 600 to 15,000 gauss. Tunable magnetrons now exist for many parts of the centimeter wave region. The tuning range for pulsed operation at high voltage is about $\pm 5\%$. It is as much as $\pm 20\%$ for low voltage magnetrons. Magnetrons may now be tuned electronically, making frequency modulation possible. Present magnetron cathodes are rugged and have long life. Even for high frequency magnetrons where current density requirements are most severe, research has made available rugged cathodes with adequate life. Magnetrons are built to withstand shock and vibration without change in characteristics. Designs have been compressed and in some cases the magnet has been incorporated in the magnetron structure in the interest of light weight for airborne radar equipments.

PART I of this paper is a general discussion of present knowledge concerning the magnetron oscillator. As such it is largely a discussion of what has come to be common knowledge among those who have carried out wartime developments. It brings together in one place results of work done by all the magnetron research groups including that at our Laboratories. PART I supplies the background necessary to understanding the

discussion in PART II of the magnetrons developed at the Bell Laboratories during the war. More complete presentations of the experimental and theoretical work done on the magnetron during the war are soon to be published by other research groups.

The material written up during the war has appeared as secret or confidential reports issued by the British Committee on Valve Development (CVD Magnetron Reports), by the Radiation Laboratories at the Massachusetts Institute of Technology and at Columbia University, and by the participating industrial laboratories. No attempt has been made in PART I to indicate the specific sources of the work done since 1940. To fit the wartime development of magnetrons into the sequence of previous developments, specific references are made to publications appearing in the literature prior to 1940.

The nature and scope of PART II of the paper are discussed more fully in its introductory Section, 11. GENERAL REMARKS.

PART I

THE MAGNETRON OSCILLATOR

1. GENERAL DESCRIPTION

1.1 *Description:* The multicavity magnetron oscillator has three principal component parts: an electron interaction space, a multiple resonator system, and an output circuit. Each of these is illustrated schematically in Fig. 1. The electron interaction space is the region of cylindrical symmetry between the cathode and the multisegment anode. In this region electrons emitted from the cylindrical cathode move under the action of the DC radial electric field, the DC axial magnetic field, and the RF field set up by the resonator system between the anode segments. These electronic motions result in a net transfer of energy from the DC electric field to the RF field. The RF interaction field is the fringing electric field appearing between the anode segments, built up and maintained by the multicavity resonator in the anode block. RF energy fed into the resonator system by the electrons is delivered through the output circuit to the useful load. The output circuit shown in Fig. 1 consists of a loop, inductively coupled to one of the hole and slot cavities, feeding a coaxial line.

To operate such a magnetron oscillator, one must place it in a magnetic field of suitable strength and apply a voltage of proper magnitude to its cathode, driving the cathode negative with respect to the anode. This voltage may be constant or pulsed. In the latter case, the voltage is applied suddenly by a so-called pulser or modulator for short intervals, usually of about one microsecond duration at a repetition rate of about 1000 pulses per second. With suitable values of the operating parameters, the magnetron

oscillates as a self excited oscillator whenever the DC voltage is applied. The energy available at the output circuit may be connected, as in a radar set, to an antenna or, as in a laboratory experimental setup, may be absorbed in a column of water.

1.2 *Analogy to Other Oscillators:* In its fundamental aspects, the magnetron oscillator is not unlike other and perhaps more familiar oscillators. In particular, instructive analogies may be drawn between the magnetron oscillator, the velocity variation oscillator, and the simple triode oscillator. In Fig. 2 is depicted schematically the parallelisms between these types of oscillators and a simplified equivalent lumped constant circuit.

In the triode of Fig. 2(a), as in the gap of the second cavity of the velocity variation tube of Fig. 2(b) and in the interaction space of the magnetron
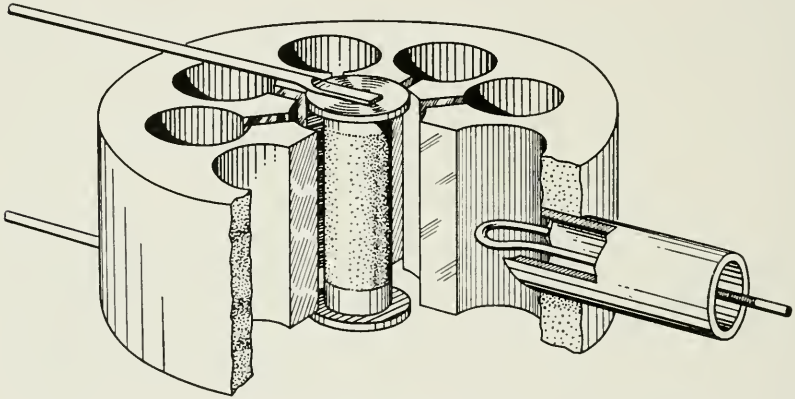
Fig. 1.—A schematic diagram designed to show the principal component parts of a centimeter wave magnetron oscillator. The resonator system and output circuit each represents one of several types used in magnetron construction.

oscillator of Fig. 2(c), electrons are driven against RF fields set up by the resonator or "tank circuit," to which they give up energy absorbed from the primary DC source.   In each type of oscillator there is operative a mechanism of "bunching" which allows electrons to interact with the RF field primarily when the interaction will result in energy transfer *to* the RF field. In the triode oscillator this is accomplished by grid, whose RF potential is supplied by the "tank circuit" in proper phase with respect to the RF potential on the anode.   In the velocity variation oscillator, bunching is accomplished by variation of the electron velocities in the gap of the first cavity, followed by drift through the intervening space to the second gap. The first cavity is driven in proper phase by a feed back line from the second cavity.   In the magnetron oscillator, as is to be described in detail later, electron interaction with the RF fields is such as to group the electrons into
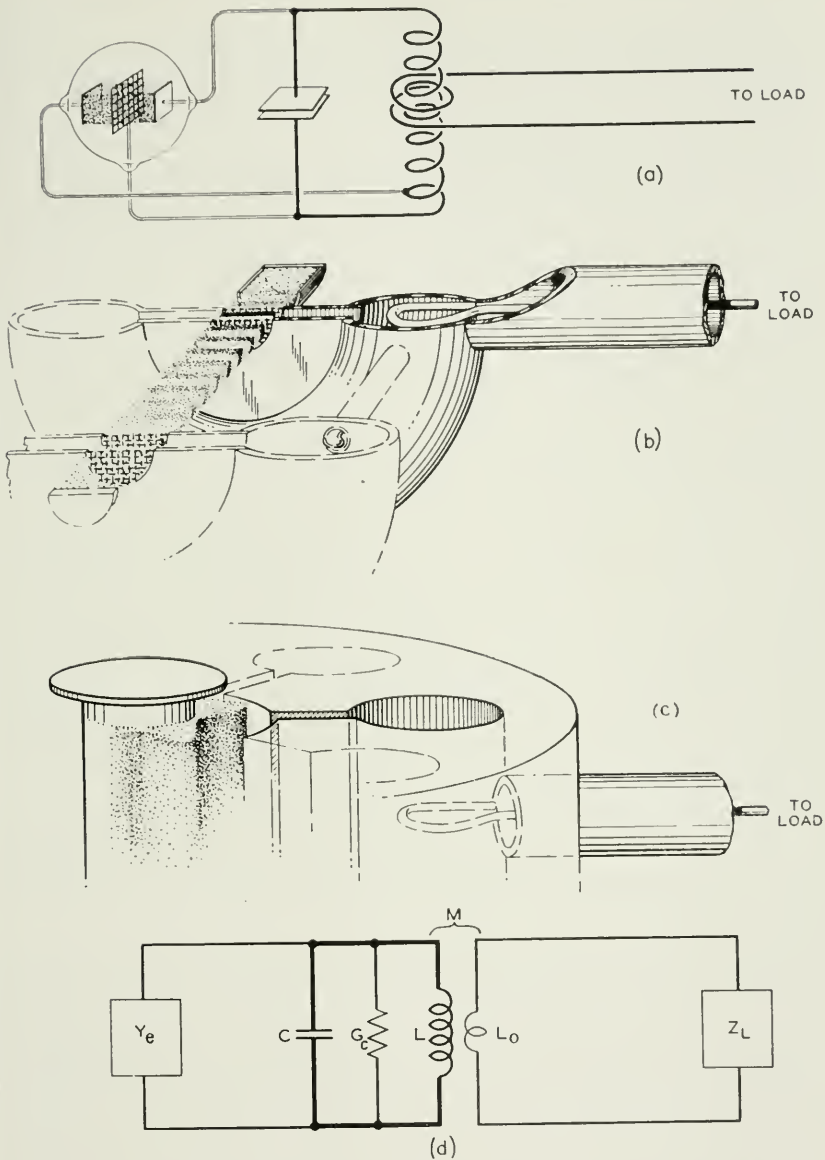
Fig. 2.—A schematic diagram depicting the parallelism among the conventional triode oscillator, the velocity variation oscillator, the centimeter wave magnetron and an equivalent lumped constant circuit. In the figure an attempt is made to align corresponding parts vertically above one another.

bunches or "spokes" which sweep past the gaps in the anode, in phase for favorable interaction with the RF fields across the gaps.   The bunching field

is thus the same field as that to which energy is transferred.   In this sense the magnetron oscillator is perhaps more properly analogous to the reflex type of velocity variation oscillator, in which a single cavity is used both as "buncher" and "catcher"; the electrons, after traversing the gap once, are turned back in the proper phase in the drift space so as to pass through the gap again in the opposite direction.

Each type of oscillator has a resonator in which energy is stored and which synchronizes the flow of energy from the electrons into it by the means of self excitation.   In each type, energy is extracted from the resonator by an output circuit at a rate which, under steady state conditions, equals that of influx from the electron interaction, minus the losses in the resonator itself.

1.3 *Use of Equivalent Circuits:* In many instances the understanding of electromagnetic oscillators is made easier and analytic treatment made possible by use of an equivalent circuit with lumped constants.   Of several possible types, one of the simpler and more frequently used for the magnetron oscillator is shown in Fig. 2.   This may appear in the case of the multi-cavity magnetron to be an oversimplification as it does not account for the fact that the resonant frequency of the magnetron resonator system is many valued.   A magnetron resonator, being made up of a number of coupled resonating cavities, is capable of supporting several modes of oscillation. These modes of oscillation have different resonant frequencies and correspond to different configurations of the electromagnetic fields.   By means to be discussed, however, magnetron resonators can be made to oscillate "cleanly" in one of these modes and may thus be represented for many purposes by a simple L-C circuit having a single resonance.

The output circuit of the oscillator is also amenable to treatment by equivalent lumped constant circuits which account for its behavior with accuracy.   More general, four terminal network theory has also been applied in the study and design of impedance transformations in this part of the oscillator.

Finally, the electrons, which in a sense are connected to the circuit formed by the resonator and the load, may also be treated by circuit concepts. The electrons moving in the space between the cathode and anode, by virtue of their presence and motion, induce charge fluctuations on the anode segments.   The time derivative of these fluctuations is equivalent to an RF current flowing into the anode from the interaction space.   This current and the RF voltage on the anode, bearing a definite phase relationship, make possible the definition of an admittance called the average electronic admittance, $Y_e = G_e + jB_e$.   Since the electrons are being driven against RF fields in the interaction space, this admittance looking into the electron stream has a negative conductance term.   Unlike usual circuit admittances, the electronic admittance is nonlinear, being a function of the voltage

amplitude of oscillation, V_{RF}, as well as of other parameters governing the electronic behavior of the oscillator, such as voltage and magnetic field. It is not known *a priori* but may be deduced from measurements on the operating oscillator and its circuit.

A necessary condition for oscillation, applicable to the magnetron as to any oscillator, is that, on breaking the circuit at any point, the sum of the admittances looking in the two directions is zero. Thus, if the circuit is broken at the junction of electrons and resonator, as is convenient, the electronic admittance, Y_e, looking from the circuit into the electron stream, must be the negative of the circuit admittance, Y_s, looking from the electron stream into the circuit.

With these remarks, of general applicability to all types of electromagnetic oscillators, the discussion will be continued for the centimeter wave magnetron oscillator in particular. As far as is possible, the electronic interaction space, resonator system, and output circuit of the device will be taken up in that order. The function and operation of each part will be described; then the principles of its design, and its relation to previous magnetron development will be indicated.

1.4 *Electron Motions in Electric and Magnetic Fields—The DC Magnetron:* Before beginning the discussion of the electronics of the magnetron oscillator, it would be well to review briefly electron motions in various types and combinations of electric and magnetic fields, and the operation of the DC magnetron.[1]

An electron, of charge e and mass m, moving in an electric field of strength E, is acted upon by a force, independent of the electron's velocity, of strength eE, directed oppositely to the conventional direction of the field. If the field is constant and uniform, the motion of the electron is identical to that of a body moving in a uniform gravitational field.

An electron moving in a magnetic field of strength B, however, is acted upon by a force which depends on the magnitude of the electron's velocity, v, on the strength of the field, and on how the direction of motion is oriented with respect to the direction of the field. The force is directed normal to the plane of the velocity and magnetic field vectors and is of magnitude proportional to the velocity, the magnetic field, and the sine of the angle, θ, between them. Thus the force is the cross or vector product of $\overrightarrow{v}$ and $\overrightarrow{B}$,

$$\overrightarrow{F} = e[\overrightarrow{v} \times \overrightarrow{B}], \qquad F = Bev \sin \theta.$$

An electron moving parallel to a magnetic field ($\sin \theta = 0$) feels no force. One moving perpendicular to a uniform magnetic field ($\sin \theta = 1$) is con-

---

[1] The cylindrical DC magnetron was reported by A. W. Hull. Phys. Rev. 18, 31 (1921).

strained to move in a circle by the magnetic force at right angles to its path. Since this force is balanced by the centrifugal force, the radius, ρ, of the circular path depends on the electron's momentum and the strength of the field; that is,

$$Bev = \frac{mv^2}{\rho},$$

yielding
$$\rho = \frac{mv}{eB}. \tag{1}$$
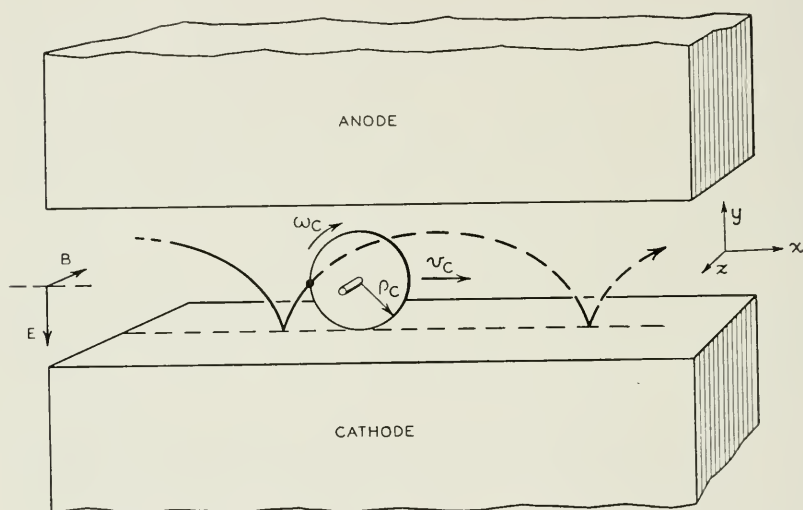
Fig. 3.—The cycloidal path of an electron which started from rest at the cathode in crossed electric and magnetic fields for the case of parallel plane electrodes. The mechanism of generation of the orbit by a point on the periphery of a rolling circle is depicted.

The time, T_c, required to traverse the circle is independent of the radius and hence of the velocity of the electron; $T_c = 2\pi\rho/v = 2\pi m/eB$. Thus, the angular frequency of traversing the circular path, the so-called cyclotron frequency, depends on the magnetic field alone and is given by,

$$\omega_c = 2\pi f_c = 2\pi \frac{1}{T_c} = \frac{e}{m}B. \tag{2}$$

In the magnetron, electron motion in crossed electric and magnetic fields is involved. Consider first such motion between two parallel plane electrodes, neglecting space charge. If, as in Fig. 3, the electric field is directed in the negative y direction and the magnetic field in the negative z direction, the equations of motion of the electron are:

$$\frac{d^2x}{dt^2} = \frac{eB}{m}\frac{dy}{dt},$$

$$\frac{d^2y}{dt^2} = \frac{e}{m}\left(E - B\frac{dx}{dt}\right),$$ (3)

$$\frac{d^2z}{dt^2} = 0.$$

The particular case of most interest here is that for which the electron starts from rest at the origin. The equations of motion then yield a cycloidal orbit given by the parametric equations:

$$x = v_c t - \rho_c \sin \omega_c t = \rho_c(\omega_c t - \sin \omega_c t),$$

$$y = \rho_c(1 - \cos \omega_c t),$$ (4)

in which:

$$v_c = \frac{E}{B},$$ (5)

$$\rho_c = \frac{m}{e}\frac{E}{B^2},$$ (6)

$$\omega_c = \frac{e}{m}B.$$ (7)

This motion may be regarded as a combination of rectilinear motion of velocity v_c in the direction of the x axis, perpendicular to both E and B, and of motion in the xy plane about a circular path of radius ρ_c, at an angular frequency ω_c, the cyclotron frequency. Fig. 3 shows the resulting cycloidal path and its generation by a point on the periphery of the rolling circle. In the case of cylindrical geometry, it is often convenient to think in terms of the plane case.

In the case of cylindrical geometry with radial electric and axial magnetic fields, the electron orbit, neglecting space charge, approximates an epicycloid generated by rolling a circle around on the cylindrical cathode. The orbit is not exactly an epicycloid because the radial motion is not simple harmonic, which state of affairs arises from the variation of the DC electric field with radius. The approximation of the epicycloid to the actual path is a convenient one, however, because the radius of the rolling circle, its angular frequency of rotation, and the velocity of its center, for the epicycloid, all approximate those for the cycloid of the plane case. These approximations improve with increasing ratio of cathode to anode radii. Several electron orbits in a DC cylindrical magnetron are shown in Fig. 4 for several magnetic fields.

It is clear from this simplified picture of the orbits in a DC cylindrical magnetron without space charge, that, at a given electric field, an electron orbit for a sufficiently strong magnetic field may miss the anode completely and return to the cathode.   The critical magnetic field at which this is just possible is called the cut-off value, B_c.   For a given voltage between cathode and anode, as the magnetic field is increased, the current normally passed by the device falls rather abruptly at B_c.   A current versus magnetic field curve, together with electron orbits corresponding to four regions of the
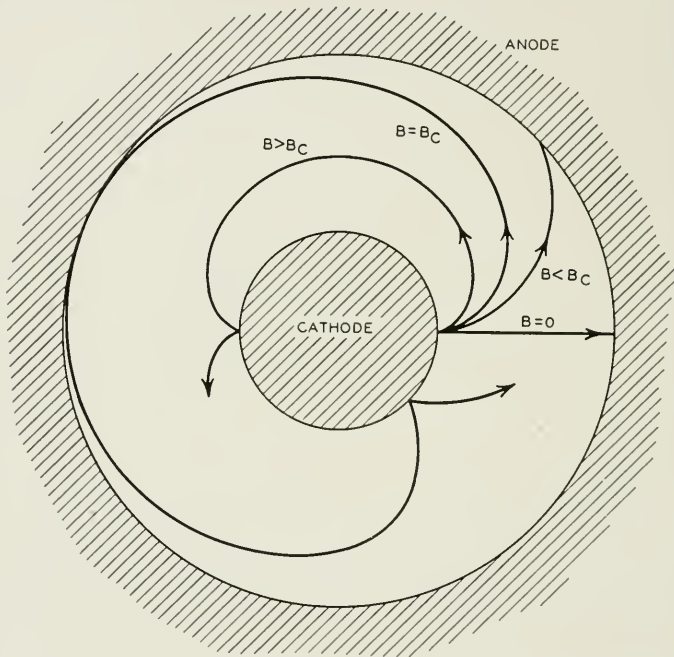
Fig. 4.—Electron paths in a cylindrical DC magnetron at several magnetic fields above and below the cut-off value, B_c.   The electrons are assumed to be emitted from the cathode with zero initial velocity.

curve, are shown in Fig. 5.   For the case of parallel plane electrodes, the cut-off relation between the critical anode potential, V_c, and magnetic field, B_c, and the electrode separation, d, for the parallel plane case, is obtained by equating the electrode separation to the diameter of the rolling circle. Thus,

$$d = 2 \frac{m}{e} \left(\frac{V_c}{d} \right) \frac{1}{B_c^2},$$

from which

$$V_c = \frac{e B_c^2 d^2}{2m}.$$

For the cylindrical case, the relation may be shown to be

$$V_c = \frac{eB_c^2 r_a^2}{8m}\left[1 - \left(\frac{r_c}{r_a}\right)^2\right]^2,\qquad(8)$$

in terms of cathode and anode radii, r_c and r_a.

2. Types of Magnetron Oscillators

2.1 *Definitions:* The DC magnetron may be converted into an oscillator, suitable for the generation of centimeter waves, by introducing RF fields into the anode-cathode region. This may be done by applying between anode
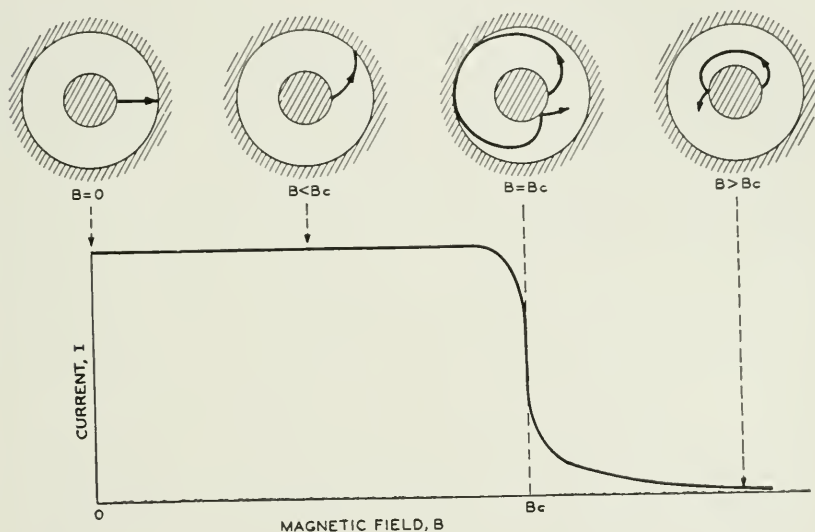
Fig. 5.—Variation of current passed by a cylindrical DC magnetron at constant voltage, plotted as a function of magnetic field. The orbits of electrons occurring at four different magnetic fields are shown above the corresponding regions of the current characteristic.

and cathode RF voltage from a resonant circuit, in which case the electrons interact with the superposed radial RF field. Or, it may be done by splitting the magnetron anode into two or more segments between which the RF voltage is applied. Then the electrons interact with the fringing RF fields existing between the segments. The problem of understanding the electronics of the multicavity magnetron oscillator is that of understanding how an electron, subject to the constraints placed upon its motion by the DC axial magnetic and DC radial electric fields, can move so as to interact favorably with the RF field; how an electron interacting unfavorably is rejected; and why, on the average, the electrons transfer more energy to the RF field than they take from it.

On the basis of the nature of the electronic mechanism by means of which energy is transferred to the RF field, it is now convenient to distinguish three types of magnetron oscillators.[2] The *negative resistance magnetron oscillator* depends on the existence of a static negative resistance characteristic between the two halves of a split anode.[3] The *cyclotron frequency magnetron oscillator* operates by virtue of resonance between the period of RF oscillation and the period of the cycloidal motion of the electrons (rolling circle or cyclotron frequency).[4] The *traveling wave magnetron oscillator* depends upon resonance, that is, approximate equality, between the mean translational velocity of the electrons and the velocity of a traveling wave component of the RF interaction field.[5]

The magnetron oscillator with which this paper is primarily concerned is of the traveling wave type. The other magnetron types are discussed briefly for the sake of completeness and because an understanding of them enhances one's grasp of the entire subject and places the traveling wave magnetron oscillator in its proper historical perspective.

2.2   *The Negative Resistance Magnetron Oscillator—Type I:* In the negative resistance magnetron oscillator,[6] the anode is split parallel to the axis into two halves, between which the RF circuit is attached. The electrons emitted by the cathode must move under the combined action of the DC radial electric and DC axial magnetic fields together with the RF electric field existing between the two semicylinders forming the anode. The transit time from cathode to anode is not involved in the mechanism except that it must be small relative to the period of the RF oscillation.   The static negative resistance characteristic arises from the fact that under certain conditions the allowable orbits for the majority of electrons terminate on the segment of lower potential, irrespective of the segment toward which they start. These electrons, being driven against the RF component of the field, give energy gained from the DC field to the RF field.

In Fig. 6 are shown the paths, plotted by Kilgore,[7] of two electrons starting initially toward opposite segments but both striking the segment of lower potential.   Each path is completely traversed in a time during which

[2] The magnetron oscillator discussed by Hull, in which the magnet winding is coupled to the plate circuit, is not considered as it is essentially an audio frequency device.   K. Okabe in his book, "Magnetron-Oscillations of Ultra Short Wavelengths" (Shokendo, 1937), distinguishes five types, but it is not clear just how his types C and E are to be identified.

[3] These oscillations have been called Habann, quasi-stationary, or dynatron oscillations, and correspond to Okabe's type D.

[4] These oscillations have been called electronic oscillations by Megaw, transit time oscillations of the first order by Herriger and Hülster, and correspond to Okabe's type A.

[5] These oscillations are the running wave type discussed by Posthumus, the transit time oscillations of higher order of Herriger and Hülster, and correspond to Okabe's type B.

[6] This type was disclosed by Habann, Zeit f. Hochfrequenz. 24, 115 and 135 (1924).

[7] G. R. Kilgore, Proc. I.R.E. 24, 1140 (1936).

the RF field changes little. Thus it is possible by applying DC potential differences between the anode segments to measure a negative resistance between them. As can be seen from the orbits of Fig. 6, magnetic fields considerably above the cut-off value are used. With magnetrons of this type, power output up to 100 watts at 600 mc/s at an efficiency of 25% has been attained.[7] Oscillations of frequency as high as 1000 mc/s, (30 cm.) have been produced.[8] Because a large number of orbital loops are required, however, making $\omega \ll \dfrac{eB}{m}$, this type of magnetron oscillator demands the
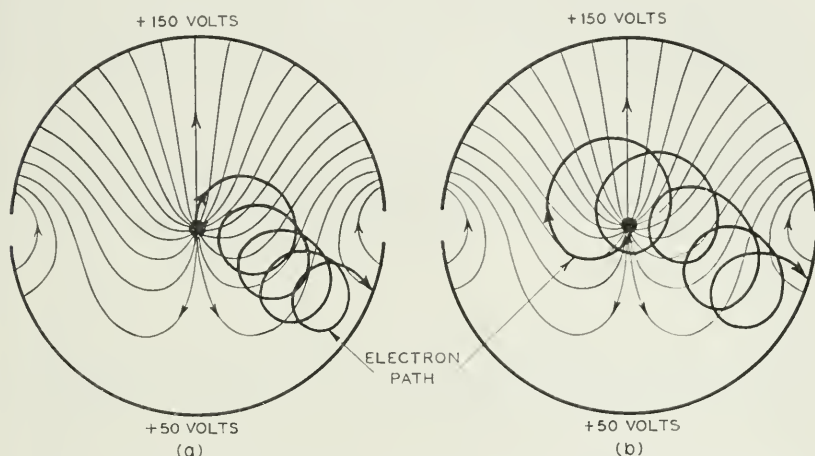
Fig. 6.—Electron paths plotted by Kilgore for the negative resistance magnetron oscillator, Type I. During the time the orbits shown are being executed, the cathode is at zero potential and the anode segments at the potentials indicated. Lines of electric force on an electron are plotted in this figure. The two orbits are those of electrons which start initially toward opposite anode segments. It should be noted that in either case the electron is driven to the segment of lower potential against the RF field component.

use of high magnetic field in the centimeter wave region and is thus less desirable than other types.

2.3 *The Cyclotron Frequency Magnetron Oscillator—Type II:* Not long after the invention of the DC magnetron, oscillations between anode and cathode were found to occur near the cut-off value of magnetic field.[9] These were found to be strongest for wavelengths obeying a relation of the form:

$$\lambda = \frac{\text{constant}}{B}. \tag{9}$$

[8] E. C. S. Megaw, Journ. I.E.E. (London) 72, 326 (1933).
[9] A. Zacek, Cos. Pro. Pest. Math. a Fys. (Prague) 53, 578 (1924). A summary appeared in Zeit. f. Hochfrequenz. 32, 172 (1928).

Later, it was shown that the oscillation period is equal to the electron transit time from the vicinity of the cathode to the vicinity of the anode and back. This made it possible to calculate a value for the constant in the above equation in good agreement with experiment.[10]   The oscillation frequency is that of the rotational component of the electronic motion, that is, approximately the cyclotron frequency of equation (7).

The mechanism must be explained in terms of electrons moving in the DC radial electric and axial magnetic fields and the superposed RF radial electric field.   This may be done as follows: An electron leaving the cathode in such phase as to gain energy when moving from the cathode toward the anode will also gain energy during its return, striking the cathode with more energy than it had when it left.   There, such an electron is stopped from further motion during which it would continue to absorb energy from the
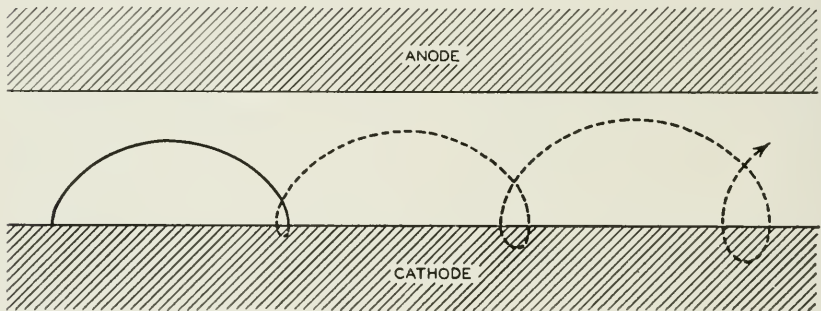
Fig. 7.—An approximate orbit of an electron which gains energy from the RF field in a cyclotron frequency or Type II magnetron oscillator, shown for the plane case.   The orbit is continued as a dashed line indicating how it would be traversed were it not stopped by the cathode.   The DC electric force on the electron is directed from cathode to anode.

RF field at the expense of the oscillation.   The electron will execute an orbit something like that of Fig. 7 for the plane case.   An electron leaving the cathode in the opposite phase, on the other hand, loses energy when moving toward the anode and again on its return toward the cathode.   As is shown in Fig. 8, it reverses its direction after the first trip without reaching the cathode surface and starts over on a second loop of smaller amplitude, remaining in the same phase and continuing to lose energy to the field.   This process continues until all the energy of the rotational component of the electron's motion has been absorbed by the RF field.   If the electron is not removed at this stage, in its subsequent motion the rotational component will build up, extracting energy from the RF oscillation.   Means such as tilting the magnetic field or placing electrodes at the ends of the tube have been used to remove the electrons from the interaction space when all the

[10] K. Okabe, Proc. I.R.E. 17, 652 (1929).

rotational energy has been absorbed. It is possible to maintain the oscillations and extract energy from them because electrons which give energy to the field can do so over many cycles, whereas electrons of opposite phase can gain energy over only one cycle before they are removed.

Magnetrons oscillating in this manner have been built with split anodes.[10,11] Here the RF field with which the electron interacts is more tangential than radial but the criterion for oscillation is the same, namely, resonance between the field variations and the rotational component of the electron's motion. Operating efficiencies of 10 to 15% have been obtained. It was with a magnetron of this type having an anode diameter of 0.38 mm. that radiation of wavelength as low as 0.64 cm. was generated.[12]

The cyclotron frequency magnetron oscillator has been almost entirely superseded by the traveling wave magnetron oscillator as a generator of
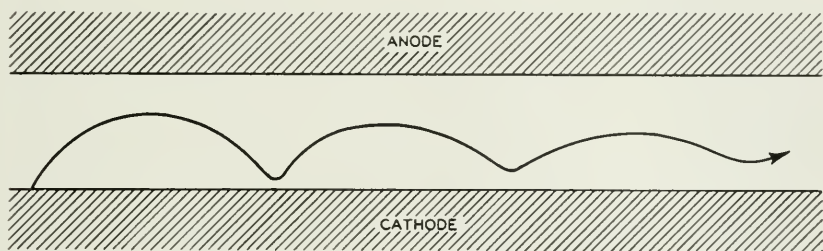
Fig. 8.—An approximate orbit of an electron which loses energy to the RF field in a cyclotron frequency or Type II magnetron oscillator, shown for the plane case. If the electron after losing all its rotational energy remains in the interaction space, it gains energy from the RF field, and its orbit builds up cycloidal scallops in a manner directly the reverse of that shown here. The DC electric force on the electron is directed from cathode to anode.

centimeter waves. In the main this is the result of the impossibility of removing electrons emitted from an extended cathode area from the interaction region at the proper stage in their orbits. This inherent drawback is not shared by the traveling wave magnetron oscillator which may be operated at higher efficiency without critical adjustment of orientation in the magnetic field or of the potential of auxiliary electrodes.

2.4 *The Traveling Wave Magnetron Oscillator—Type III:* Oscillations have been found to occur in the magnetron which are independent of any static negative resistance characteristic and which can occur at frequencies widely different from the cyclotron frequency. In 1935[13] the electronic mechanism of these oscillations was correctly interpreted as an inter-

[11] H. Yagi, Proc. I.R.E. 16, 715 (1928).
[12] C. E. Cleeton and N. H. Williams, Phys. Rev. 50, 1091 (1936).
[13] K. Posthumus, Wireless Engineer 12, 126 (1935).

action of the electrons with the tangential component of a traveling wave whose velocity is approximately equal to the mean translational velocity of the electrons. Later[14] the role of the radial component of the rotating electric field in keeping the electrons in proper phase was recognized. Magnetrons of wavelength as short as 75 cm., operating at better than 50% efficiency, were built prior to 1940, but performance such as was later to be attained with this type of magnetron at much shorter wavelengths was not attained then, perhaps primarily because of the lack of a good resonator. It was a magnetron of this type which the British brought to the United States in 1940. The British magnetron was a 10 cm. oscillator, intended for pulsed operation, having a tank circuit consisting of eight resonators built into the anode block as shown in Fig. 1.[15]

3. The Electronic Mechanism

3.1   *Electronic Interaction at Anode Gaps:* The electrons in the interaction space of the magnetron oscillator are the agents which transfer energy from the DC field to the RF field.   As such, they must move subject to the constraints imposed by the DC radial electric and DC axial magnetic fields, considering, for the moment, the RF fields to be small.   Under these conditions, as has been seen for the DC cylindrical magnetron (see Fig. 4 for $B > B_c$), electrons follow approximately epicycloidal paths which progress around the cathode.   The mean velocity of this progression, that of the center of the rolling circle, depends upon the relative strengths of the electric and magnetic fields [see equation (5) for the plane case].   By proper choice of DC voltage, V, between cathode and anode and of magnetic field, B, the mean angular velocity of the electrons may be set at any desired value.

The RF electric fields in the interaction space, with which the electrons moving as described above must interact, are the electric fields fringing from the slots in the anode surface.   These fields are provided by the N coupled oscillating cavities of which the magnetron resonator system is composed. As will be discussed in more detail later, such a system of resonators may oscillate in a number of different modes in each of which the oscillations in adjacent resonators, and thus the fields appearing across adjacent anode slots, bear a definite phase relationship.   For a system of N resonators it will be seen that the phase difference between adjacent resonators may assume the values $n\dfrac{2\pi}{N}$ radians, n being the integers 0, 1, 2, $\cdots$, $\dfrac{N}{2}$.

Adopting another point of view, one may consider the potentials placed upon the anode segments by the resonators.   The variation of the potential

[14] F. Herriger and F. Hülster, Zeit. f. Hochfrequenz, 49, 123 (1937).

[15] The use of such internal resonators is reported in the literature by N. T. Alekseieff and D. E. Maliaroff, Journ. of Tech. Phys. (U.S.S.R.) 10, 1297 (1940); republished in English, Proc. I.R.E. 32, 136 (1944).   A. L. Samuel has obtained U. S. Patent # 2,063,342 Dec. 8, 1936, for a similar device.

from one segment to the next depends upon the mode of oscillation of the system as a whole. The restriction on the phase difference stated above requires the sequence of anode segment potentials at any instant to contain
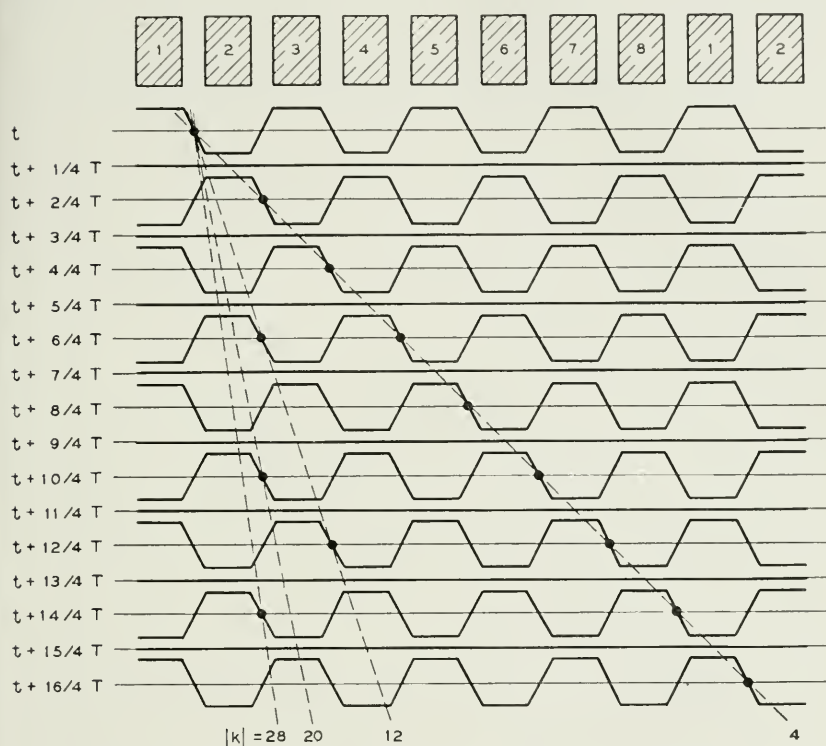
Fig. 9.—A plot showing the π mode anode potential wave at several instants in an eight resonator magnetron and the mean paths of electrons which interact favorably with the RF field. The plot is developed from the cylindrical case, the shaded rectangles at the top representing the anode segments. The anode potential variation is a standing wave, shown here for a sequence of instants one quarter period ($T/4$) apart. Note that the potential is constant across the anode surfaces and varies linearly between them. Electrons interacting favorably with the RF field cross the anode gaps when the field there is maximum retarding as indicated by the filled circles. The lines for $|k| = 4, 12, 20, 28, \cdots$ represent mean paths of electrons traveling with mean angular velocities $\frac{2\pi f}{4}$, $\frac{2\pi f}{12}$, $\frac{2\pi f}{20}$, $\frac{2\pi f}{28}$, $\cdots$ around in the interaction space. Since the field is a standing wave, it is clear that electrons possessing these velocities in either direction may interact favorably with the RF field.

n complete cycles in one traversal of the cylindrical anode, n still denoting the integers $0, 1, 2, \cdots, \frac{N}{2}$. In general, these anode potential waves may be standing waves or waves traveling around the anode structure in either direction with angular velocity $\frac{2\pi f}{n}$ radians per second, where f is the RF

frequency. For the two modes in which adjacent resonators are in phase $(n = 0)$ and π radians out of phase $(n = \dfrac{N}{2}$, the so-called π mode), however, only standing potential waves on the anode are possible. As examples of
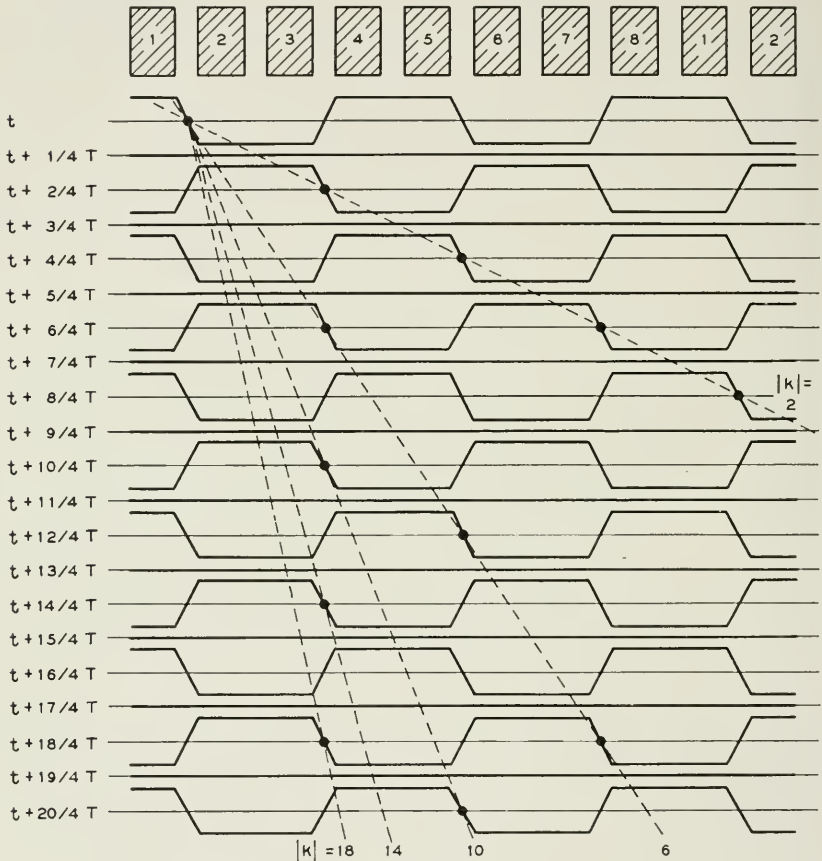
Fig. 10.—A plot similar to that of Fig. 9 for the standing wave component of anode potential of periodicity $n = 2$ in a magnetron having eight resonators. Electrons which interact favorably have mean angular velocities $\dfrac{2\pi f}{2}$, $\dfrac{2\pi f}{6}$, $\dfrac{2\pi f}{10}$, $\dfrac{2\pi f}{14}$, $\cdots$ in either direction in the interaction space.

standing and traveling anode potential waves in an anode structure having eight resonators $(N = 8)$, the standing wave for $n = 4$ and standing and traveling waves for $n = 2$ are shown in Figs. 9, 10, and 11 respectively.

From what has been said about the RF field and the electronic motion in the interaction space of the magnetron oscillator of Type III, it is possible

to understand its fundamental electronic mechanism.   As in any oscillator, the criterion for oscillation is that more energy shall be transferred to the RF field by electrons driven against it than is taken from the RF field by electrons accelerated by it.   This can be accomplished in the traveling wave
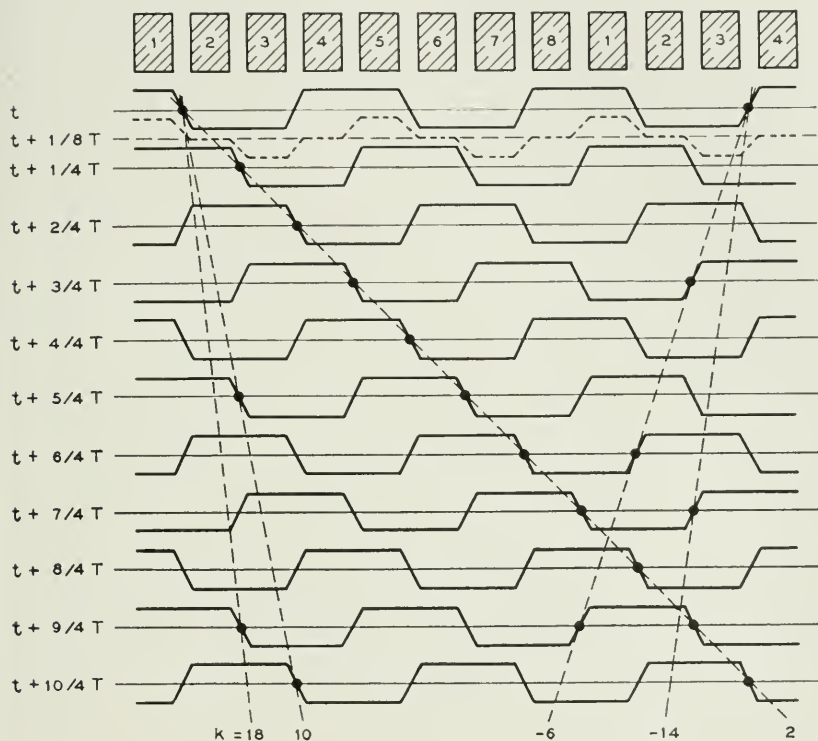
Fig. 11.—A plot similar to those of Figs. 9 and 10 for the rotating wave of anode potential of periodicity $n = 2$ in a magnetron having eight resonators [see equation (13) in the text]. The field at the instant $t + \frac{1}{8}T$ is plotted as a dashed line to show that the traveling wave does not preserve its shape at all instants.   Whereas the wave travels in one direction with the angular velocity $\frac{2\pi f}{2}$, electrons which travel with velocities $\frac{2\pi f}{2}, \frac{2\pi f}{10}, \frac{2\pi f}{18}, \cdots$ in the same direction or with velocities $\frac{2\pi f}{6}, \frac{2\pi f}{14}, \cdots$ in the opposite direction interact favorably with the RF field.   Directions of electron motion must now be distinguished.   Electrons whose velocity is opposed to that of the rotating field are said to be driving a "reverse" mode.

magnetron oscillator only if the mean angular velocity of the electrons is such as to make them pass successive gaps in the anode at very nearly the same phase in the cycle of the RF field across the gaps.   Then it is possible for an electron which leaves the cathode in such phase as to oppose the tangential component of the RF field across one anode gap, to continue to

lose energy gained from the DC field to the RF field at successive gaps. Electrons which gain energy from the RF field are driven back into the cathode after only one orbital loop and are removed from further motion detrimental to the oscillation. This process of selection and rejection of electrons forms the groups of bunches, shown in Fig. 2(c), which sweep past the anode slots in phase to be retarded by the RF field component. The criterion that the electron drift velocity shall be such as to keep these bunches in proper phase is analogous to the condition that the drift angle in a velocity variation oscillator [Fig. 2(b)] be such as to cause the bunches to cross the gap of the second or "catcher" cavity in phase to lose energy to the RF field across the gap.

The condition placed upon the mean angular velocity of the electrons may be discussed more readily by reference to Figs. 9, 10, and 11. Consider first, however, only Fig. 9 for the standing potential wave of the $n = 4$ mode, and focus attention on an electron which crosses the gap between anode segments 1 and 2 at the instant t when the RF field is maximum retarding, that is, the potential on segment 1 is maximum and on segment 2 minimum. It is clear that this electron can cross the next gap in the same phase if the time required to reach it is $(|p| + \frac{1}{2}) T$, in which p is any integer and T is the period of RF oscillation. In Fig. 9, four lines are drawn representing the mean paths of electrons moving with such velocities as to make $p = 0, 1, 2,$ and 3. Each line crosses a gap when the RF field is maximum retarding, that is, when the potential has the maximum negative slope at the center of the gap. As will be seen later, a more convenient parameter, to be called k, is that whose absolute magnitude, $|k|$, specifies the number of RF cycles required for the electron to move once around the interaction space. $\frac{|k|}{N}$ is then the number of cycles between crossings of successive anode gaps, which for the π mode of Fig. 9 must take on the values:

$$\frac{|k|}{N} = |p| + \frac{1}{2}, \qquad p = 0, \pm 1, \pm 2, \cdots,$$

or the values given by the more general expression, applicable to any mode:

$$\frac{|k|}{N} = |p| + \frac{n}{N}, \qquad p = 0, \pm 1, \pm 2, \cdots.$$

In this expression, $\frac{n}{N}$ is the phase difference between adjacent resonators, expressed as a fraction of a cycle. k may thus assume the values given by

$$k = n + pN,$$
$$p = 0, \pm 1, \pm 2, \cdots. \tag{10}$$

The mean angular velocity which the electrons must possess is then given by

$$\frac{d\theta}{dt} = \frac{2\pi}{kT} = \frac{2\pi f}{k},$$ (11)

in which θ is the azimuthal angle.

For the π mode ($n = \frac{N}{2}$) it is seen that the negative integers, p, give the same series of values for $|k|$ as do the positive integers including zero. The sequence is $|k| = 4, 12, 20, 28, \cdots$. Reference to Fig. 9 indicates that electrons may travel in either direction around the interaction space and interact favorably with the RF field, provided their mean angular velocity is given by equation (11) with values of k specified by equation (10). That this should be so is clear from the fact that the anode potential wave is a standing wave with respect to which direction has no meaning. Fig. 9 also makes clear how an electron moving with velocity different from that corresponding to the lines shown, will fall out of step with the field, and on the average be accelerated as much as it is retarded, thus effecting no net energy transfer.

In Figs. 10 and 11, diagrams for the $n = 2$ mode similar to that of Fig. 9 for the π mode are shown. Fig. 10 is for electronic interaction with a standing wave of periodicity $n = 2$ and Fig. 11 for a traveling wave of the same periodicity. Again, as in the case of the π mode, the values of k for favorable electronic interaction are given by equation (10).

The sequence of positive integral values of p (including zero) and the sequence of negative integral values of p do not each give the same sequence of values for $|k|$ as was the case for the π mode. For $p \geqq 0$, $|k| = 2$, 10, 18, $\cdots$, and for $p < 0$, $|k| = 6, 14, 22, \cdots$. For the standing potential wave (Fig. 10) each of these values of $|k|$ does specify the velocity of possible electron motion in *either* direction for favorable interaction with the field. For the traveling potential wave (Fig. 11), on the other hand, only the positive values of k ($p \geqq 0$) correspond to electron motion in the same direction as the traveling wave, the negative values of k ($p < 0$) corresponding to electron motion in the direction opposite to the traveling wave. The sign of k has significance. If the electrons are moving with velocities specified by equation (11) with the negative values of k from equation (10), and are thus moving counter to the traveling RF field wave, the electrons are said to be driving a "reverse" mode.

The actual electron orbits do not correspond to simple translation but, as has been discussed, to rotation superposed on translation. The epicycloid-like scallops in the orbit are of no significance to the fundamental electronic mechanism. It is the mean velocity of the electron motion around the interaction space, specified by the relative values of V and B, that is of importance. The magnetron may operate, for example, at such

high magnetic field, provided V has the proper value, that the scallops became relatively small variations in an otherwise smooth orbit (see Fig. 18).

In the cylindrical magnetron, the radial variation of the DC electric field, resulting in a decrease in the mean angular velocity of the electrons as they approach the anode, would make it impossible for an electron to keep step with the fields across the anode gaps were not a mechanism of phase focusing operative. That such focusing is inherent in the interaction of electrons and fields will be seen later.

3.2 *The Interaction Field:* The electronic mechanism which has been discussed in terms of electron motions through the fields at the gaps in the multisegment anode, may also be discussed in terms of the traveling waves of which the RF interaction field may be considered to be composed. The RF interaction field corresponds to anode potential waves like those plotted in Figs. 9, 10, and 11. The interaction fields for the several modes of oscillation of the resonator system are thus to be distinguished by the number, n, of repeats of the field pattern around the interaction space. Since the potential at the anode radius is nearly constant across the faces of the anode segments and varies primarily across the slots, the azimuthal variation of the field cannot be purely sinusoidal but must involve higher order harmonics.

For a mode of angular frequency $\omega = 2\pi f$, corresponding to a phase difference between adjacent resonators of $n\frac{2\pi}{N}$, the anode potential wave is of periodicity n around the anode and may be written as a Fourier series of component waves traveling in opposite directions around the interaction space:

$$V_{RF} = \sum_k A_k e^{j(\omega t - k\theta + \gamma)} + \sum_k B_k e^{j(\omega t + k\theta + \delta)},$$

$$k = n + pN, \qquad p = 0, \pm 1, \pm 2, \cdots. \tag{12}$$

Note that the summations are taken over all integral values of k given by equation (10).

The interaction field for any mode of periodicity n is thus represented by two oppositely traveling waves, whose fundamentals are moving with angular velocities $\frac{\omega}{n} = \frac{2\pi f}{n}$, and whose component amplitudes, A_k and B_k, in general are not equal. γ and δ are arbitrary phase constants. The expression (12) may be reduced to the form:

$$V_{RF} = \sum_k (A_k - B_k) \cos(\omega t - k\theta + \gamma)$$

$$+ \sum_k 2B_k \cos\left(\omega t + \frac{\gamma + \delta}{2}\right) \cos\left(k\theta - \frac{\gamma - \delta}{2}\right), \tag{13}$$

$$k = n + pN, \qquad p = 0, \pm 1, \pm 2, \cdots,$$

which shows that the complete field pattern may be considered to consist of a rotating wave superposed on a standing wave, each having a fundamental component of periodicity n.

The fact that the periodicities, k, of the harmonics in the expressions (12) or (13) are those for which k has the values given by (10) may be determined from a Fourier analysis of the complete anode potential waves like those of Figs. 9, 10, and 11. Only those harmonics which specify the same pattern of potentials at the centers of the anode segments as the fundamental are admitted in the analysis.

As has been mentioned before, the complete field patterns for $n = 0$ and $n = \dfrac{N}{2}$ are standing waves. Thus for these modes of oscillation $A_k = B_k$ in the expressions (12) and (13). For the other modes, $n = 1, 2, 3, \cdots$, $\dfrac{N}{2} - 1$, the electrons may interact with the traveling field component of expression (13) or with the standing field components which, in case $A_k = B_k$, is the only component present (see Figs. 10 and 11 for the case $n = 2$, $N = 8$).

The terms in expressions (12) and (13) for which $|k| = n$ are the fundamental components; those for which $|k| \neq n$ are called the Hartree harmonics. The components of field strength corresponding to these harmonics in the interaction field pattern fall off in intensity from anode to cathode more rapidly the higher the value of k. The variation with radius is of the form $\left(\dfrac{r}{r_a}\right)^k$. Thus the farther from the anode one samples the field, the more like the fundamental sinusoidal pattern it appears.

For each value of k in expression (12), whether or not $A_k = B_k$, there are two oppositely traveling sinusoidal wave components of periodicity k. Since each requires k cycles of the RF oscillation to complete one trip around the interaction space, the linear velocity at the anode surface is $\dfrac{2\pi f r_a}{k}$ corresponding to an angular velocity of $\dfrac{2\pi f}{k}$. Thus, as seen in Fig. 11 for the instant $t + T/8$, the change of shape of the total traveling wave indicates that the components of which it is composed travel with different velocities. In Fig. 23 are shown instantaneous RF interaction field patterns for the fundamental components ($p = 0$) of the $n = 1, 2, 3$, and 4 modes of an anode block having eight resonators.

3.3   *The Traveling Wave Picture:* It is instructive to discuss the operation of the Type III magnetron oscillator in terms of electron interaction with the traveling wave components present in the interaction field. This might at first appear to be difficult in view of the many components of several possible modes. By mode frequency separation, as discussed later, it is

generally possible, however, to restrict oscillation to only one mode, usually the π mode. Further, the fact that the electronic motion in crossed DC electric and magnetic fields results in a mean drift of electrons around the interaction space, enables one to restrict his attention to a single traveling wave corresponding to the fundamental or a single Hartree harmonic of the field of this mode; for it is possible, in principle at least, by proper adjustment of V and B to equate the mean angular velocity of the electrons to the angular velocity, $\dfrac{2\pi f}{|k|}$, of *any* one of the traveling field components. When this is true, only the field of this component has an appreciable effect upon the electron's motion. With respect to the fields of the oppositely traveling component of the same harmonic (same k), and the components of all other harmonics (different k), the electron finds itself drifting rapidly through regions of accelerating and decelerating field with no net energy transfer. From the point of view of the electron, the fields of the other components vary so rapidly as to average out over any appreciable interval of time. The only exception to these statements occurs when a harmonic of periodicity k' of another mode of frequency f' has the same angular velocity as the harmonic of periodicity k, that is, when $\dfrac{2\pi f'}{|k'|} = \dfrac{2\pi f}{|k|}$. The effect on magnetron operation of this coincidence of angular velocities will be discussed further in a later section. In the calculation of electron motions, the restriction to the field of a single traveling wave component has been called the rotating field approximation.

The consideration of the electronic mechanism has thus been reduced to that of the motion of electrons under the combined influence of the radial DC electric field, the axial DC magnetic field, and a sinusoidal field wave traveling around the interaction space. From what has been said thus far it is clear that for energy to be transferred to the RF field it is necessary that the mean electron velocity very nearly equal that of the traveling wave. Then an electron leaving the cathode in such phase as to find itself moving in a region of decelerating tangential component of the RF field, may continue to move with this region and lose energy to the field. In contrast to the Type II magnetron oscillator, the energy transferred to the RF field in this case is the potential energy of the electron in the radial DC electric field. The energy in the rotational component of the motion remains practically unaffected and the electron orbit from cathode to anode looks something like that plotted in Fig. 12 for the case with plane electrodes. On the other hand, an electron which leaves the cathode in such phase as to gain energy in a region of accelerating tangential RF field, is removed at the cathode after only one cycle of the epicycloid-like motion. If this did not occur, the electron would continue to move with the field and absorb energy. An approxi-

mate orbit is shown in Fig. 13. It is instructive to compare the orbits of the two categories of electrons in the traveling wave magnetron oscillator with the orbits of corresponding electrons in the cyclotron frequency type of magnetron oscillator (Figs. 7 and 8). In each case, it is the fact that "favorable" electrons may interact for a considerably longer time than "unfa-
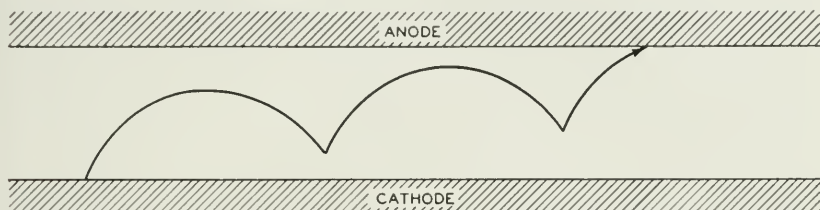
Fig. 12.—An approximate orbit of an electron which is losing energy to the RF field in a traveling wave or Type III magnetron oscillator, shown for the plane case. Here the energy loss is potential energy of the electron in the DC field. Compare this with the orbit in Fig. 8 where the energy loss is rotational energy of the electron. The DC electric force on the electron is directed from cathode to anode.
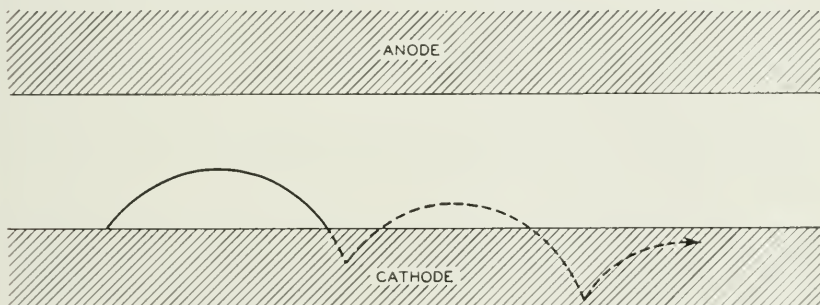
Fig. 13.—An approximate orbit of an electron which gains energy from the RF field in a traveling wave or Type III magnetron oscillator, shown for the plane case. The orbit is extended as a dashed line as though the cathode were not there. The energy gained is potential energy of the electron in the DC field. Compare this with the orbit in Fig. 7 where the energy increase is in the rotational energy of the electron. The DC electric force on the electron is directed from cathode to anode.

vorable" electrons which makes possible a net energy transfer between the DC and RF fields.

One may now compare the traveling wave picture of the electronic mechanism with that presented earlier in which the motion of electrons past the gaps in the anode structure is considered. An electron moving so that $\frac{|k|}{N} = |p| + \frac{n}{N}$ cycles of the RF oscillation elapses between its crossing of two successive anode gaps, is thus moving around the interaction space in synchronism with a traveling component of the kth harmonic of the inter-

action field.   Both points of view are of value.   That involving the motion of electrons past the anode gaps is more fundamental physically.   That in terms of a traveling wave component, on the other hand, is more convenient in calculations of electron orbits including space charge effects, where by transformation to a coordinate system rotating with the field it is possible to treat of motions in static fields.

3.4   *Phase Focusing:* It has been seen from two points of view how groups of electrons which move around the interaction space of the magnetron oscillator are formed by a process of selection and rejection of electrons by the tangential component of the RF field.   However, space charge de-bunching and the discrepancy at all but one radius between the mean veloc-ity of translation of the electrons and the velocity of the interaction field would tend to disperse these groups and prevent efficient interaction, were it not for the phase focusing provided by the radial component of the RF field.

The mechanism of the phase focusing may be discussed either in terms of the interaction of electrons with the actual fields existing at the anode gaps or in terms of the traveling wave picture of the electronic mechanism.   The fundamental mechanism involved depends upon the effect of the radial component of the RF field in aiding or opposing the radial DC field in de-termining the mean drift velocity of the electron around the interaction space.   If the radial RF field increases the net radial field in which the electron finds itself at any instant, the mean velocity of the electron in-creases as can be seen from equation (5) for the plane case.   Similarly, a decrease in the net radial electric field, caused by the RF radial component, results in decreased electron translation velocity.   These changes in the electron's velocity operate so as to keep the electron near the position in which it can interact most favorably with the tangential component of the RF field.

Consider an electron which crosses an anode gap at the instant the RF field there is maximum retarding, that is, an electron which is to be found on the plane marked M in Fig. 14 at this instant.   It experiences about as great an increase of velocity by virtue of the radial component aiding the DC radial field before crossing the gap as decrease by virtue of the radial component opposing the DC radial field after crossing the gap.   Another electron which is lagging behind the electron just considered is to be found opposite a positively charged anode segment, as at P in Fig. 14, when the RF field passes through its maximum value.   Since the RF field component decreases with time after this instant, the effect of the radial component of the field on the electron's velocity after crossing the gap will be less than its effect before crossing the gap, the net effect being one of increasing the mean velocity of translation, bringing the electron more nearly into the proper phase.   An electron which leads the electron first considered, on the other

hand, will be found opposite the negatively charged anode segment beyond the gap when the RF field is maximum, and for it the net effect of the radial component is to reduce the mean velocity of the electron, bringing it also more nearly into the proper phase.

In discussing the mechanism of phase focusing from the traveling wave point of view, the field lines of Fig. 14 may be considered to be those of the traveling wave component with which the electrons are interacting. Then the whole field pattern indicated moves to the right as shown by the arrow above the plane of maximum retarding tangential field at M. An electron
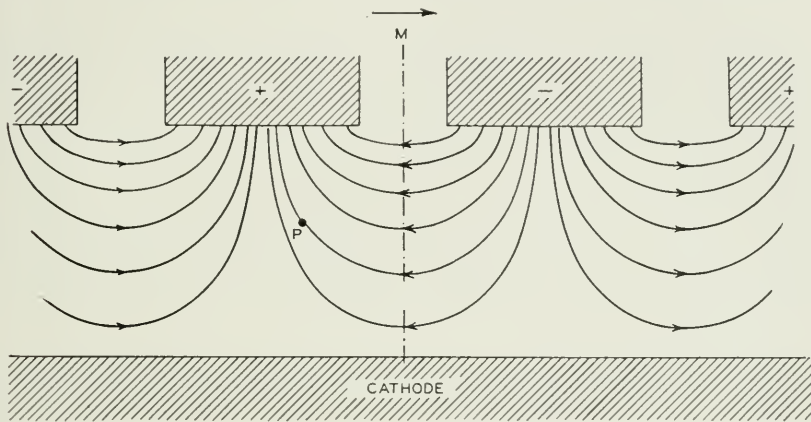
Fig. 14.—A plot of lines of electric force on an electron (drawn for the plane case) for the fundamental of the π mode. It is shown for the purpose of explaining the phase focusing property of the radial field component. The plane of maximum opposing force on the electron intersects that of the figure along the line M. The arrow shown above the line M indicates the direction of electron motion. The force on the electron due to the DC electric field is directed from cathode to anode. The force lines shown may be considered to be those of the total fundamental field component at the instant the field is maximum. Then an electron at the point P will cross the center of the anode gap after the instant of maximum retarding force.   Or the lines shown may be considered to be those of the traveling component of the fundamental moving in the same direction as the electrons. In this case an electron at the point P is lagging behind the maximum of the retarding tangential field.

which falls behind the position M to the point P, for example, finds itself in a stronger net radial electric field which increases its mean translational velocity tending to bring it back to the position M. The reverse holds for an electron which runs ahead of the plane M.

3.5 *Space Charge Configuration:* The over-all picture of the electronic mechanism in the Type III magnetron oscillator thus presents a spoke-shaped space charge cloud of electrons wheeling around the cathode in synchronism with the anode potential wave, each spoke in a region of maximum retarding field. This picture of what is happening has been very handsomely confirmed by actual orbital calculations taking account of

space charge. The calculations have been carried out by the so-called self consistent field method using the rotating field approximation mentioned earlier. In this method, orbits of electrons are calculated in an assumed field, and the space charge due to these electrons determined. The field calculated on the basis of this space charge distribution may then be compared with that assumed. This cycle of calculations is repeated, each time using the calculated field of the previous cycle as that in which the electrons move, until a field is obtained which is consistent with that used in calculating the electron orbits. This method will be recognized as that used in the calculation of electron orbits about the nucleus in atoms.

The result of one such calculation is shown in Fig. 15. The orbits of four electrons which were emitted from the cathode in different phases are plotted in a set of coordinates rotating with the RF field component. One electron is returned to the cathode, and the other three reach the anode. The boundaries of the space charge cloud are shown as dashed lines. The spoke-shaped structure is clear, and its position with respect to the rotating anode potential wave is as expected. The number of spokes of the cloud is equal to the periodicity of the component of the mode with which the electrons are interacting. In the case of Fig. 15 there are four spokes, since the magnetron is operating in the fundamental of the $n = 4$ mode ($k = 4$, $p = 0$).

3.6 *Induction by the Space Charge Cloud:* Another view of the mechanism by which the electrons drive the resonator system may be obtained by considering the effect of the space charge spokes in inducing current flow in the anode segments themselves. For example, the oscillation of the resonator block in its π mode corresponds to the periodic interchange of electric charge from each anode segment, around a resonating cavity to the next anode segment. This oscillation is maintained, much in the manner of a pendulum escapement drive, by the space charge spoke appearing in front of an anode segment at that instant in the oscillation cycle when it can aid in building up the net positive charge on the segment. At the same instant, the adjacent segments, being opposite a "gap" in the space charge wheel, may build up a negative charge.

The RF current, I_{RF}, induced in the anode structure, thus results from the motion of the spoke-shaped space charge cloud in the interaction space. It is not to be confused with the total circulating RF current in the resonator system. Whereas I_{RF} must be in a phase with the space charge cloud it need not be in phase with the RF voltage, V_{RF}, between the anode segments. In terms of the electron motions, this means that the spokes of the space charge cloud may lead or lag the maxima in the tangential field. In general, the electronic admittance defined by the ratio of I_{RF} to V_{RF} may thus include a susceptance as well as a conductance. The product of V_{RF} and the in-phase component of I_{RF}, integrated over a period of one cycle of RF

oscillation, equals the energy per cycle which is delivered to the load. This amount of energy is twice that transferred in the half cycle during which the spokes of space charge move against the field from positions in front of one
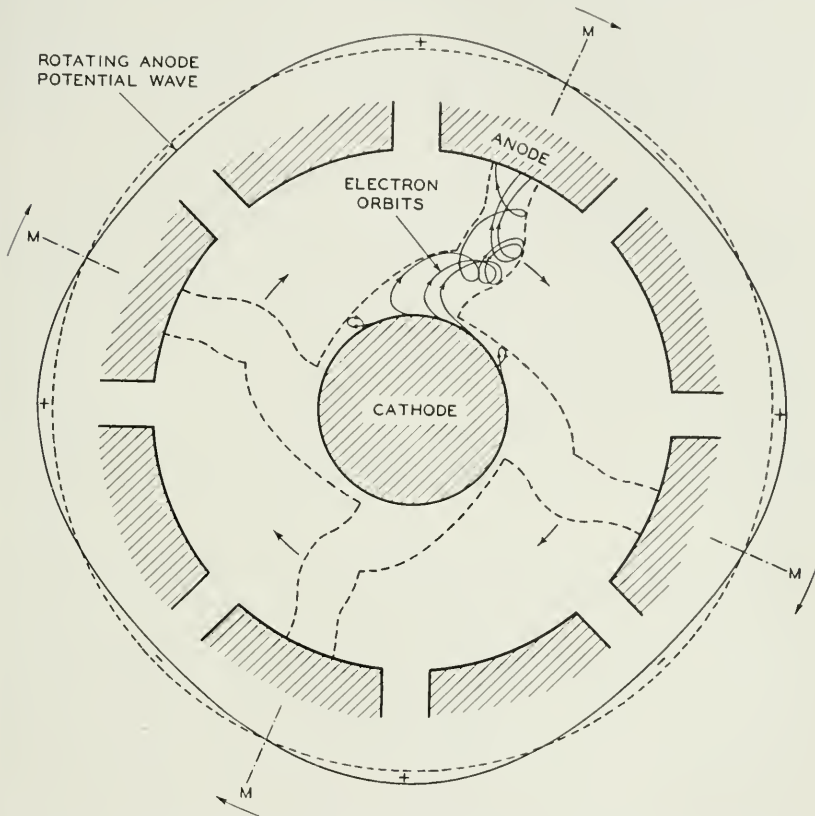
Fig. 15.—The orbits of four electrons which left the cathode in different phases, plotted n a coordinate system rotating with the anode potential wave. These orbits have been icalculated by the self consistent field method which includes space charge effects. In this frame of reference the orbits exhibit loops whereas in a stationary frame they would more nearly exhibit cusps. The dashed lines inclose the orbits of the electrons and hence delineate the boundaries of the space charge cloud which rotates around the cathode in synchronism with the anode potential wave. Planes of maximum retarding tangential field are represented by the lines *M* (see Fig. 14). This figure is reproduced by courtesy of the British Committee on Valve Development (CVD) and is taken from the CVD Magnetron Report No. 41.

set of alternate anode segments to similar positions in front of the adjacent anode segments.

In each spoke of the electron space charge cloud, individual electrons progress from cathode to anode. The DC current, I, passed by the magnetron is made up of electrons which strike the anode from the ends of the space charge spokes. Quite apart from its dependence on other parameters,

this DC current is directly proportional to anode length, h. If the magnetron is driven at greater DC current, the space charge in the interaction space increases but the phase of its structure with respect to the traveling anode wave does not change to a first approximation. Thus both the in-phase and quadrature components of I_{RF} increase with no change in electronic admittance. The second order effects which do arise from small shifts in the phase of the rotating space charge structure are discussed in Section 10.4 *Electronic Effects on Frequency.*

4. Conditions Relating Measurable Parameters

4.1 *Necessary Conditions for Oscillation:* After having discussed the electron motions in the interaction space of the Type III magnetron oscillator, the viewpoint will now be changed to that looking from the outside in, so to speak, and it will be asked what conditions relating measurable parameters are imposed by the nature of the electronic mechanism. Beyond the geometrical parameters of cathode and anode radii, r_c and r_a, one can determine the DC voltage, V, applied between cathode and anode; the magnetic field B, in which the magnetron is placed; the DC current, I, drawn by the anode; the frequency of oscillation, f; and, from impedance measurements, the RF load, $Y_s = G_s + jB_s$, presented to the electrons by the resonator, output, and load.

Perhaps the most fundamental condition for oscillation of the traveling wave magnetron is that imposed by the requirement of synchronism between the electron drift and the RF field. As has been indicated, the angular velocity of a rotating component of a Hartree harmonic of the interaction field, of order k, is $\dfrac{2\pi f}{|k|}$. An approximate expression for the mean angular velocity of the electrons may be determined by neglecting the variation of electric field with radius and calculating the angular velocity midway between cathode and anode, thus:

$$\frac{v}{(r_a + r_c)/2} = \frac{E/B}{(r_a + r_c)/2} = \frac{V/(r_a - r_c)B}{(r_a + r_c)/2} = \frac{2V}{(r_a^2 - r_c^2)B}.$$

Equating this to the angular velocity $\dfrac{2\pi f}{|k|}$, one obtains the relation:

$$V = \frac{\pi f}{|k|} r_a^2 B \left[1 - \left(\frac{r_c}{r_a}\right)^2 \right]. \tag{14}$$

In this derivation it should be recognized that the angular velocity $\dfrac{2\pi f}{|k|}$ may be considered either to be that of a traveling component of the RF field with which the electron interacts or the mean angular velocity which the

electron must have to maintain proper phase with the total RF fields existing across the anode gaps.

Posthumus[13] derived an expression, assuming negligible cathode diameter, which is similar to equation (14). By the same method as that used above, Slater has derived an expression differing from (14) by a term which results from the use of a more accurate value for the electron's translational velocity at the midpoint between cathode and anode in cylindrical geometry. Slater's expression is:

$$V = \frac{\pi f}{|k|} r_a^2 B \left[1 - \left(\frac{r_c}{r_a} \right)^2 \right] - 2 \frac{m}{e} \left(\frac{\pi f r_a}{|k|} \right)^2 \left[1 - \left(\frac{r_c}{r_a} \right)^2 \right]. \quad (15)$$

Hartree has derived an expression from a consideration of the conditions under which electrons are just able to reach the anode with infinitesimal amplitude of RF voltage in the kth harmonic. It is:

$$V = \frac{\pi f}{|k|} r_a^2 B \left[1 - \left(\frac{r_c}{r_a} \right)^2 \right] - 2 \frac{m}{e} \left(\frac{\pi f r_a}{|k|} \right)^2. \quad (16)$$

In a sense this condition represents a cut-off relation for the oscillating magnetron analogous to Hull's cut-off relation for the DC magnetron [equation (8)].

Plotted on a V-B graph, the expressions (14), (15), and (16) represent parallel straight lines. The line of equation (14) passes through the origin; the so-called Hartree line of (16) is tangent to the DC cut-off parabola; the so-called Slater line of (15) lies above the Hartree line but below the line of expression (14). Each of the above expressions indicates that the electrons will drive a given harmonic of the RF interaction field in a type III magnetron oscillator only at values of DC voltage and magnetic field which bear a definite relation. This relation expresses the fact that V/B is very nearly constant [equation (14)].

In Fig. 16 are plotted as an illustration, the Hartree lines for the fundamentals ($p = 0$) of the $n = 1, 2, 3$, and 4 modes and for the $k = -5$ harmonic ($p = -1$) of the $n = 3$ mode of a 10 cm. magnetron with eight resonators.[16]

[16] This magnetron, used as an illustrative example in Part I, has the following characteristics:

Number of resonators, N = 8.
Cathode radius, r_o = 0.3 cm.
Anode radius, r_a = 0.8 cm.
Anode length, h = 2.0 cm.
Anode to end cover distance = 0.6 cm.
Frequency (π mode), f = 2800 mc/s.
Wavelength (π mode), λ = 10.7 cm.
DC operating voltage, V = 16 kv.
Operating magnetic field, B = 1600 gauss.
Typical operating peak current, I = 20 amps.
Over-all efficiency, η = 42%.
Peak power output, P_o = 135 kw.
Pulse duration, τ = 1 microsecond.
Pulse repetition frequency = 1000 pps.

Since the operating voltage is found to increase with increasing current, oscillation at a constant anode current takes place along a line, such as the Slater line for example, lying slightly above and parallel to the Hartree line (see Fig. 16). The separation of the operating line from the Hartree line increases with increasing DC current.
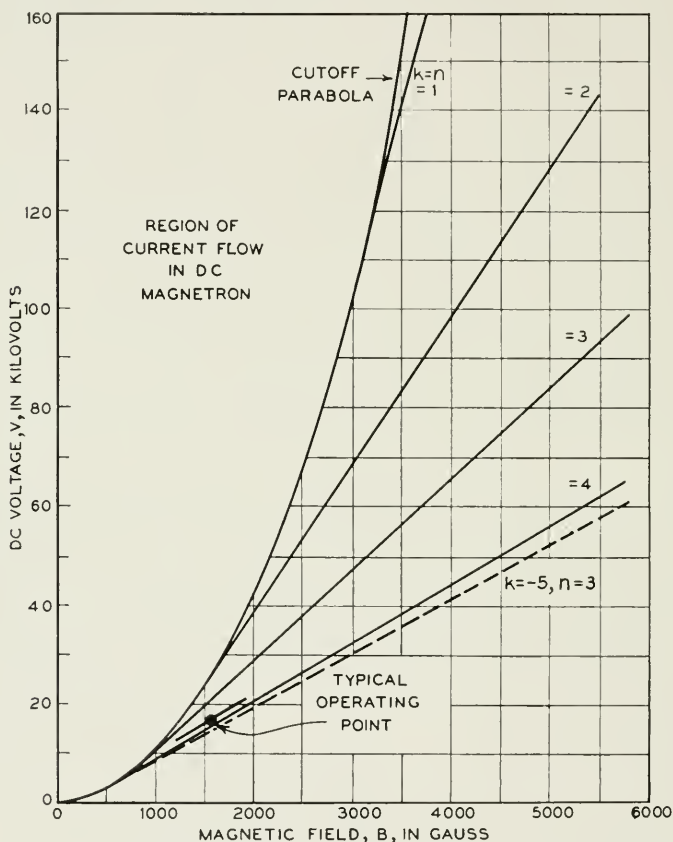
Fig. 16.—A *V-B* plot for a magnetron having eight resonators (see footnote 16) showing the cut-off parabola and Hartree lines for several rotating field components. The ranges of DC voltage and magnetic field have been extended considerably beyond values ever applied to such a magnetron to show the lines for the fundamentals of all of the modes. The typical operating point plotted is that specified in footnote 16 and plotted in Fig. 17.

The necessary conditions for oscillation discussed above have been of great value in the identification of the modes of operating magnetrons and as the starting point in the design of new magnetrons for given wavelength, magnetic field, and voltage.

4.2 *The Performance Chart:* Another fundamental performance char-

acteristic of the operating magnetron is the *V-I* plot or performance chart.
In Fig. 17 such a chart is plotted for the same magnetron[16] used as the ex-
ample for Fig. 16. In it are plotted contours of constant magnetic field,
RF power output, and over-all efficiency. The fact that the constant mag-
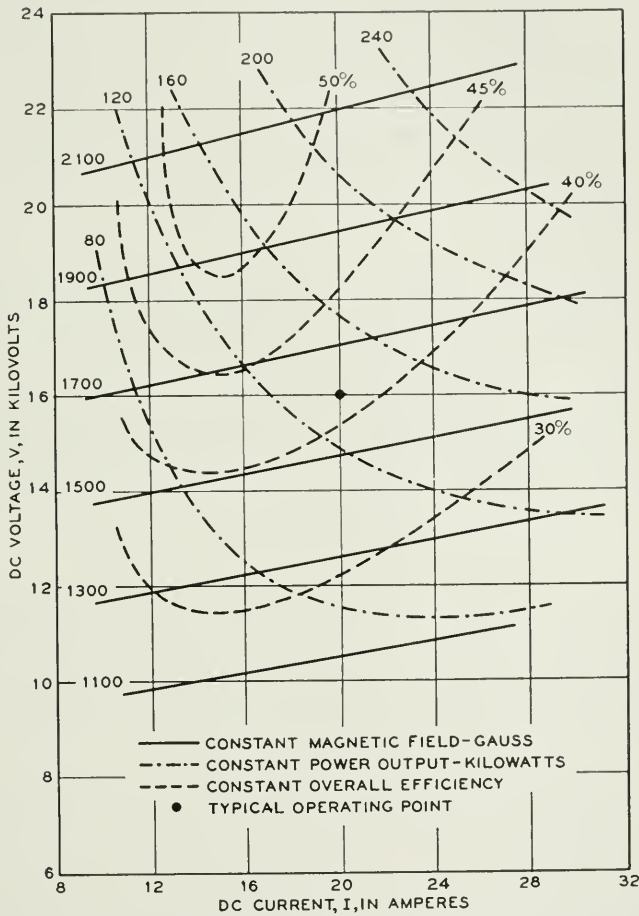
Fig. 17.—A *V-I* plot or performance chart for a magnetron having eight resonators
(see footnote 16). Contours of constant magnetic field, power output, and over-all effi-
ciency are shown. The typical operating point plotted is that specified in footnote 16 and
plotted in Fig. 16.

netic field contours are nearly horizontal and spaced as they are is a mani-
festation of the oscillation conditions of equations (15) and (16). The in-
crease of voltage with current is an effect, attributable to the space charge,
quite independent of the condition of synchronism between field and elec-

trons.   For if the magnetron is to deliver more power at a given magnetic field, the induced RF current must increase.   This entails increased space charge and a greater DC current flow.   To maintain the increased space charge additional DC voltage is required.

4.3   *The Electronic Efficiency:*   The performance chart also shows the not too surprising fact that more power may be drawn from the magnetron as the voltage and current are increased.   More interesting are the increase of the over-all efficiency with voltage and the maximum through which the efficiency passes with increasing current.   This variation of over-all efficiency, η, is to be attributed to changes in the electronic efficiency, η_e, since the other factor involved in the over-all efficiency, the circuit efficiency, η_c, is essentially constant over the diagram $(\eta = \eta_e\eta_c)$.

Fig. 18.—Approximate orbits of electrons which transfer energy to the RF field, plotted for operation of a plane magnetron at two different magnetic fields.   It is shown how the relative kinetic energy gained beyond the last cusp and dissipated at the anode decreases as the magnetic field and voltage of operation are increased, resulting in increased efficiency of electronic conversion of energy from the DC field to the RF field.   The two illustrative cases differ by a factor two in DC voltage and hence by the same factor in magnetic field and diameter of the rolling circle.

The increase of electronic efficiency with voltage, and hence magnetic field, may be explained by the picture of electron motions in the interaction space.   The highest electronic efficiency is attained when the electrons reaching the anode do so with least kinetic energy.   For the approximate orbit of Fig. 12, the energy lost at the anode per electron is that gained as kinetic energy beyond the last cusp.   Bringing the last cusp closer to the anode, corresponding to a reduction of the amplitude of the rotational component of the electron motion, reduces the ratio of kinetic energy lost at the anode to the potential energy possessed by the electron at the cathode. Thus, according to equation (6) for the radius of the rolling circle, this fractional energy loss should vary as V/B^2 or, since V/B is approximately constant, as $1/B$, η_e increasing with B.   In terms of the approximate electron orbits for a plane magnetron, Fig. 18 shows how increase in voltage

and magnetic field increases the electronic efficiency.   The dependence of electronic efficiency on B predicted by this simple picture is in accord with the dependence predicted by more sophisticated theories.

In all probability the decrease of electronic efficiency toward low and high currents is, in part at least, the result of decrease in the phase focusing action. This occurs as a result, on the one hand, of low RF field strength in the proper mode when the current and RF oscillation are small, and, on the other hand, as a result of space charge debunching when the current and space charge become very large.   In addition, at low currents the leakage to the anode of electrons which are not effective in interaction with the RF field, assumes a more dominant role, reducing the effective electronic efficiency.   These electrons are no doubt those emitted near or at the ends of the cathode.

Of importance to the motion of electrons near the ends of the interaction space, and thus to the electron leakage, are the configurations of the DC electric and magnetic fields there.   These depend upon the geometry of the cathode ends and surrounding walls and of the magnetic pole pieces.   The electrons are largely confined to the interaction space by the axial force, directed toward the center of the interaction space, produced by the non-uniformities in the electric and/or the magnetic fields.   For uniform magnetic field, the desired focusing action on the electrons may be achieved by disks at cathode potential which are mounted at each of the cathode and extend beyond the cathode surface over the ends of the interaction space as may be seen in Fig. 1.   In other cases, distortion of the magnetic field in the end spaces of the magnetron, in addition to cathode end disks, has been used to produce the same effect.

Although the dependence of operation of a magnetron oscillator on load is primarily a circuit problem, detailed discussion of which will be delayed until the RF circuit has been discussed, there is one feature of the dependence on load and circuit characteristics which may properly be discussed now.   This is the dependence of the electronic efficiency, η_e, on the circuit conductance, G_s, presented to the electrons.   The plot in Fig. 19 is a typical example and shows an optimum value of conductance, to each side of which η_e decreases. With decreasing G_s the RF voltage increases.   Whereas initially the increase of V_{RF} with decreasing G_s increases the phase focusing properties, it results eventually in such strong RF field that electrons are drawn more or less directly to the anode where they arrive with considerable kinetic energy. On the other hand, the decrease in V_{RF} with increasing G_s eventually will result in an RF field too weak to produce the necessary amount of phase focusing.   The value of G_s presented to the electrons depends not only upon the output circuit properties and load but also upon the parameters of the resonator itself.   The dependence of electronic efficiency upon G_s is a

good example of how intimately the electronics and circuit of the magnetron oscillator are associated.

4.4 *Scaling:* Once an efficient design has been achieved for a given wavelength, voltage, current, and magnetic field, one is interested in reproducing it at other values of these parameters. In doing this, use is made of the theory of scaling. For cases where the interaction space remains geometrically similar and the magnetron operates in the same mode, the same efficiency is presumably achieved when it is arranged that the electron orbits remain similar. Directly from dimensional arguments applied to Maxwell's equations for the electromagnetic field and to the equations of
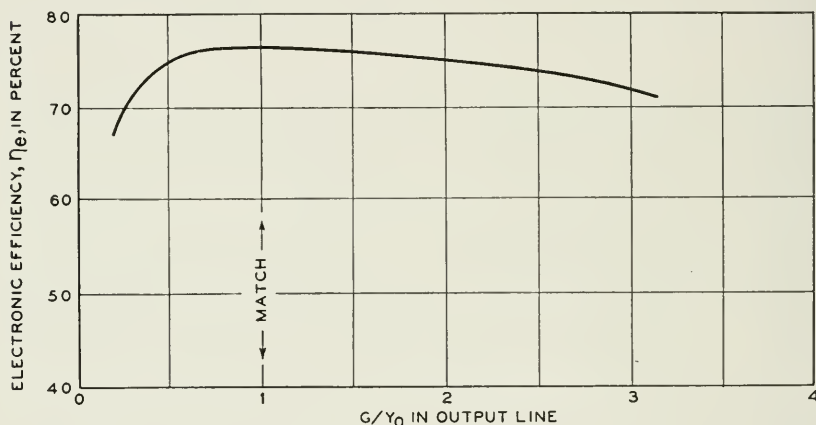
Fig. 19.—A plot of electronic efficiency as a function of load conductance. The conductance G/Y_o in the output line is related to that presented by the circuit to the electrons, G_e, directly through the magnetron resonator and output circuits. The magnetron with which these data were obtained is a 3.2 cm. magnetron having sixteen resonators.

motion of the electrons, it can be shown that the orbits and operation will be equivalent for all conditions for which the loading and the quantities

$$\lambda B, \quad (\lambda/r_a)^2 V, \quad \text{and} \quad (\lambda^3/r_a^2 h)I$$

remain invariant.

It would perhaps be of interest to consider, as a simple example, the case for which all the linear dimensions of a given magnetron are changed by a factor α. The new resonant wavelength, λ', is equal to $\alpha\lambda$ since the new resonator is α times the size of the original, while the new frequency is $f' = f/\alpha$. The new anode radius, cathode radius, and anode length scale directly so that

$$r'_a = \alpha r_a, \ r'_c = \alpha r_c \text{ and } h' = \alpha h$$

Since λB, $(\lambda/r_a)^2 V$, and $(\lambda^3/r_a^2 h)I$ must remain invariant,

$$\lambda' B' = \lambda B \text{ or } B' = B/\alpha,$$

$$(\lambda'/r'_a)^2 V' = (\lambda/r_a)^2 V \text{ or } V' = V,$$

and $$(\lambda'^3/r'^2_a h') I' = (\lambda^3/r_a^2 h) I \text{ or } I' = I.$$

Thus the magnetic field changes by $1/\alpha$ and the operating voltage and current remain unchanged.

The idea of scaling has been extended to magnetrons of varying r_c/r_a and N by the introduction of sets of reduced variables for V, B and I involving r_c/r_a and N, in terms of which the performance may be expressed independently of the exact nature of the magnetron.

4.5   *Effect of Other Components in the Interaction Field:*   In the discussion of the traveling wave magnetron oscillator, it has been considered that the electrons interact with a single traveling RF field component, generally a component of the fundamental of the π mode.   The justification for this, as has been stated, is that from the point of view of the electron the fields of all other Hartree harmonics of the π mode vary so rapidly that their effects average out over an appreciable number of RF cycles.   This is generally true, as well, for the harmonics of other modes which may be excited. The fact has been mentioned that it is possible for the values of V and B for oscillation in the π mode very nearly to satisfy equation (16) for oscillations in a harmonic of another mode.   Then the angular velocity, $\dfrac{2\pi f'}{|k'|}$, of the harmonic very nearly equals that of the π mode, $\dfrac{2\pi f}{N/2}$, and the Hartree line of the harmonic lies very close to that of the π mode (see Fig. 16).   The effect of this situation on magnetron operation will be discussed in connection with the problem of "moding" in Section 10.6 *Oscillation Buildup—Starting.*

Of particular interest is the presence in the interaction space, in addition to the π mode field, of an RF field component which is independent of angle. Such a component appears, for example, as an inherent contamination in the interaction field of the so-called "rising sun" type resonator system to be described later.   Generally, this component is of no concern because the electrons interact with the π mode component throughout a time interval covering several cycles, during which the effect of the contamination averages out.   The exception to this occurs when the frequency of the rotational component of the electron motion, approximately the cyclotron frequency, resonates with the RF frequency as in a Type II magnetron.   From equation (2) this occurs for the plane case when $f = c/\lambda = eB/2\pi m$, from which $\lambda B = 2\pi cm/e = 10{,}700$ gauss-cm.   For a typical cylindrical case the constant is somewhat greater, being about 12,500 gauss-cm.   When λB has this

value, perturbation of the electron motion occurs, decreasing the efficiency of interaction with the traveling wave and manifesting itself on the perform-
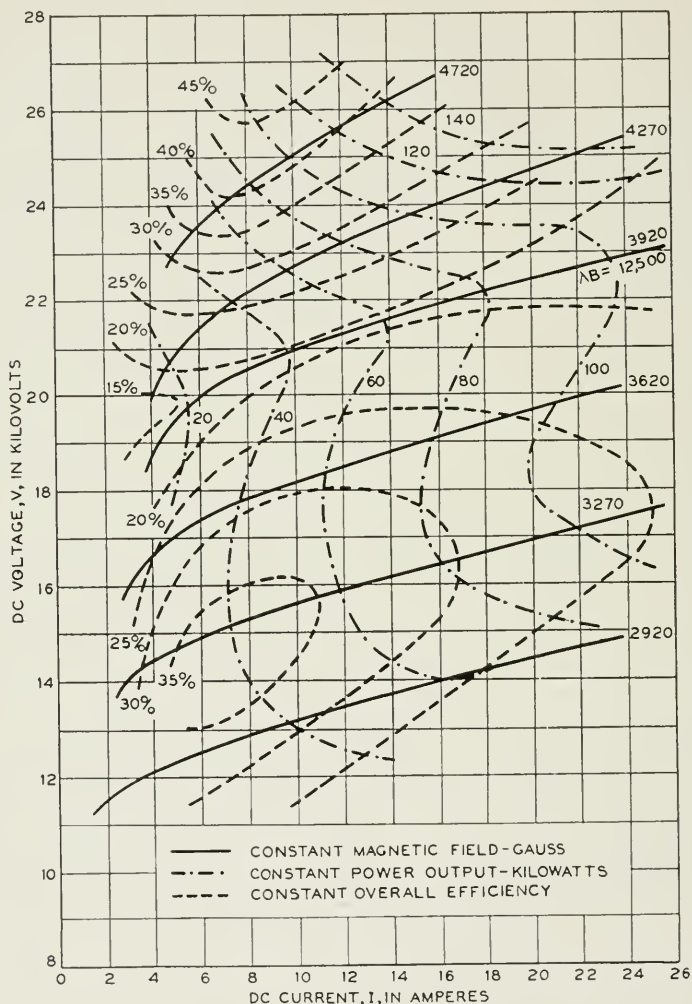
Fig. 20.—A performance chart for a magnetron which has present in its interaction field a contaminating component independent of azimuth. Note the efficiency "valley" appearing along the line $B = 3920$ gauss (compare with Fig. 17). The wavelength of the magnetron being 3.18 cm., λB along the efficiency "valley" is 12,500 gauss-cm. The data were obtained on a magnetron having a "rising sun" type resonator system and are reproduced here through the courtesy of the Columbia Radiation Laboratory.

ance chart as a "valley" in operating efficiency appearing along the constant B line for which $\lambda B = 12,500$ gauss-cm. In Fig. 20 is plotted the performance chart of a magnetron, having a "rising sun" type anode structure which

exhibits this kind of behavior.   In general, it would be preferable to operate this magnetron at magnetic fields above the "valley," but considerations having to do with available magnetic fields, in some cases may make it necessary to operate near the efficiency maximum at lower magnetic fields.

5. The RF Circuit of the Magnetron Oscillator

5.1   *General Considerations:* The discussion to this point has presented a picture of how the multicavity magnetron oscillator works from the point of view of its electronics; how in this respect it is related to other types of magnetron oscillators; and how, on the basis of the picture of the electronic mechanism, some of its fundamental operating characteristics are to be accounted for.   In the same manner the RF circuit of the magnetron oscillator, comprising the resonator system, the output circuit, and the load, will be discussed.   The importance of this part of the magnetron is apparent. It provides the RF fields with which the electrons interact.   To do this, electromagnetic energy must be stored in the cavities of the resonator, which reservoir in turn is tapped to deliver energy through the output circuit to the useful load.   The detailed manner in which these functions are performed has a bearing, not only on such circuit characteristics as circuit efficiency or on the effect of load on frequency, but, as has been seen, on the electronic efficiency as well.   Furthermore, the circuit analysis of the magnetron oscillator enables one to explain the remaining important operating characteristic, the so-called Rieke diagram, which describes the operational dependence on load.

The type of resonator system used in the magnetron oscillator of concern here has already been indicated in Fig. 1.   It is a resonator system comprising a number of cavities spaced equally about the cylindrical anode. This general shape is dictated by the slotted anode cylinder upon which the RF interaction field is set up.   To be sure, other types of resonators have been devised which contrive to place π mode potentials on the anode segments of a cylindrical magnetron.   Here, however, except for brief references, the discussion will concern itself with the multicavity resonator system of the general type shown in Fig. 1.   Although the individual cavities have not been limited to hole and slot geometry like those shown in Fig. 1, and other features have been added, a resonator system consisting of a system of cavities, arranged around the anode in the manner of Fig. 1, has been used in the majority of magnetron oscillators developed for centimeter wave generation since 1940.

5.2   *Simple Single Frequency Resonator:* The fact that the magnetron resonator system has a number of cavities electromagnetically coupled together makes it multiresonant.   What has been learned about the various modes and their electromagnetic field configurations, and how they may in a sense be controlled to improve magnetron operation must be discussed in some detail.   Before this is done, however, it would be well to refresh one's memory as to the fundamental ideas concerned with a single electromagnetic

resonator like one of the magnetron cavities.[17]   This is important not only because the magnetron resonator system comprises a number of such cavities but also because the resonator system as a whole may under certain circumstances be considered to resonate at but one frequency, in which case it behaves like a simple single frequency resonator.

Concerned with the simple electromagnetic resonator are the fundamental ideas of a natural frequency of resonance, of the rate of energy loss or the sharpness of resonance, and of the characteristic admittance or the energy storage capacity.   The electromagnetic resonator, whether it has lumped or distributed constants, consists of a device in which energy is transferred between electric and magnetic fields cyclically in a manner entirely analogous to the transfer of energy between potential and kinetic in the simple swinging pendulum.   Each of these oscillations, electromagnetic or mechanical, is described by a second order differential equation in terms of a parameter such as voltage or current on the one hand, and angular displacement of the pendelum bob on the other.   The solutions represent simple harmonic oscillations, the one for the simple electrical circuit having the frequency,

$$ f_0 = \frac{\omega_0}{2\pi} = \frac{1}{2\pi \sqrt{LC}} . \tag{17} $$

This occurs when the susceptances of the two components of the circuit, L and C, are equal, or when

$$ \frac{1}{\omega_0 L} = \omega_0 C. \tag{18} $$

The fact that a finite time is required to transfer the energy between the electric and magnetic fields is a lumped constant circuit is not surprising since a finite time is required for a condenser to charge or discharge, and for a current to build up or decay in an inductance.

5.3   *The Q Parameters:* The type of oscillation of the simple L-C circuit discussed above is its natural or free oscillation, not constrained by the ap-

---

[17] In the next sections of this paper, material has been drawn from the theory of a single resonant circuit having either lumped or distributed constants, the theory of coupled circuits, and the theory of centimeter wave transmission in coaxial lines and wave guides, treatments of which are to be found in the following representative texts:

L. Page and N. I. Adams, *Principles of Electricity*, D. Van Nostrand Co., New York (1931).

E. A. Guilleman, *Communication Networks*, Vols. I and II, John Wiley and Sons, New York (1931).

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

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

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

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

plication of any external driving force. It is the sort of oscillation the circuit would undergo were it left to itself after being excited or started initially. Such an oscillation, once started, does not continue indefinitely because the energy put into the circuit initially, dissipates itself in resistive losses in the circuit components and in a load which may be coupled electromagnetically to the circuit. The exponential rate at which the original energy is dissipated is a very important characteristic of the circuit. It is usually specified by a parameter Q, defined as 2π times the ratio of the energy stored in the circuit to the energy dissipated per cycle of the oscillation.[18] Thus a circuit always loses a certain fraction of its energy per cycle independent of how great this energy may be. In the exponential decay of oscillations in a resonator from which the drive has been removed, $Q/2\pi$ is the number of cycles of oscillation, required for the stored energy to decay to $1/e$ of its initial value. Similarly, in the buildup of oscillations in a resonator to which is fed a constant amount of energy per cycle, $Q/2\pi$ is the number of cycles required for the stored energy to buildup to $(1 - 1/e)$ of its final equilibrium value.

It is possible to define several types of Qs for a circuit, depending upon the nature of the energy dissipation being considered. If one considers only the energy lost in the resistance of the circuit components themselves, one defines the so-called unloaded Q, Q_o. If the circuit is coupled electromagnetically to a resistive load, the Q defined in terms of the energy dissipated in the load and internal resistance is called the loaded Q, Q_L. Finally, for some purposes it is convenient to consider the ratio of energy stored to that dissipated in the external load only. This defines the external Q, Q_{ext}. It is clear that both Q_L and Q_{ext} are functions of the degree of coupling between the oscillating circuit and the resistive load.

The Q parameters, however, tell one more than the rate at which energy is dissipated in a circuit oscillating at its natural frequency. The admittance of the circuit, measured as a function of the frequency of an external driving source, passes through a minimum at the natural frequency of oscillation of the circuit. The sharpness of the dip in the admittance curve is determined by the Q_L of the circuit in such a manner that the sharper the dip, the higher the Q_L. In passing through resonance the susceptance of the circuit changes sign from inductive for frequencies below the frequency of resonance, to capacitive, for frequencies above the frequency of resonance. The rate at which the susceptance varies with frequency is another measure of the sharpness of resonance and of the Q_L of the circuit.

5.4 *Energy Storage and Loss:* The remaining ideas concerned with a simple *L-C* circuit of lumped constants which should be mentioned here are the characteristic admittance of the circuit, the energy storage capacity,

---

[18] As will be seen in the subsequent discussion the factor 2π is included here so as to simplify the definition in terms of admittance.

and how these are related to each other and to the concepts already mentioned. For this purpose it is convenient to consider the circuit shown in Fig. 31 (c). Across the terminals AB is connected the L-C circuit in which the resistive losses are represented by the circuit conductance G_c. The circuit is loaded by the admittance $Y''_L = G''_L + jB''_L$.

Looking into the circuit at the terminals AB one sees the admittance:

$$Y_s = G_s + jB_c + Y''_L = G_c + \frac{1}{j\omega L} + j\omega C + Y''_L$$

which may be rewritten:

$$Y_s = G_c + j \sqrt{\frac{C}{L}} \left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega} \right) + Y''_L$$

$$\cong G_c + 2j \left(\frac{\omega - \omega_0}{\omega_0} \right) Y_{0c} + Y''_L, \tag{19}$$

where $\omega_0 = \dfrac{1}{\sqrt{LC}}$ and $Y_{0c} = \sqrt{\dfrac{C}{L}}$.

The expression $\sqrt{\dfrac{C}{L}}$, having the dimensions of an admittance, is by definition the characteristic admittance of the circuit, Y_{0c}. Its relation to the energy stored in the circuit, and through this to the various Qs defined above, may be seen from the following: Using the root mean square value of voltage, the energy stored in the circuit is CV^2_{RF}. This can be reduced by the use of the definition of the resonant frequency and by differentiation of the expression (19) for the admittance, thus:

$$CV^2_{RF} = \frac{Y_{0c}}{\omega_0} V^2_{RF} = \frac{1}{2} \left| \frac{dB_c}{d\omega} \right|_{\omega=\omega_0} \cdot V^2_{RF}. \tag{20}$$

Thus, at a given frequency ω_0, the energy stored in the circuit for unit voltage applied to it may be specified either by the characteristic admittance of the circuit, Y_{0c}, or by the rate of change of susceptance with frequency at the resonant frequency, $\left| \dfrac{dB_c}{d\omega} \right|_{\omega=\omega_0}$.

The loss of energy per cycle in the circuit itself, that is, in the shunt conductance G_c, is the power loss in the circuit divided by the frequency, $\dfrac{V^2_{RF} G_c}{\omega_0/2\pi}$. From this and equation (20) the unloaded Q is seen to be:

$$Q_0 = 2\pi \frac{\dfrac{Y_{0c}}{\omega_0} V^2_{RF}}{\dfrac{V^2_{RF} G_c}{\omega_0/2\pi}} = \frac{Y_{0c}}{G_c}. \tag{21}$$

Similarly, for the loaded and external Qs:

$$Q_L = \frac{Y_{0c}}{(G_e + G_L'')}, \tag{22}$$

$$Q_{\text{ext}} = \frac{Y_{0c}}{G_L''}. \tag{23}$$

It follows that the Qs are related thus:

$$\frac{1}{Q_L} = \frac{1}{Q_0} + \frac{1}{Q_{\text{ext}}}. \tag{24}$$

The efficiency of the circuit, defined as the fraction of the energy which reaches the useful load is then:

$$
\begin{aligned}
\eta_c &= \frac{G_L'' V_{RF}^2}{G_L'' V_{RF}^2 + G_e V_{RF}^2} = \frac{G_L''}{G_L'' + G_e} \\
&= \frac{\dfrac{Y_0}{Q_{\text{ext}}}}{\dfrac{Y_0}{Q_{\text{ext}}} + \dfrac{Y_0}{Q_0}} = \frac{1}{1 + \dfrac{Q_{\text{ext}}}{Q_0}} = \frac{Q_L}{Q_{\text{ext}}}.
\end{aligned} \tag{25}
$$

5.5 *Resonators with Distributed Constants.* The individual cavities of the magnetron oscillator, however, are circuits in which the parameters are distributed and not lumped. They may be considered to be "strip-type" resonators, three forms of which generally used in magnetrons, are shown in Fig. 21 (a), (b), and (c). Type (a) has been called the slot type resonator; (b), the vane type, deriving its name from a common method of fabrication in which rectangular plates are disposed around and brazed to the inside of a cylindrical cavity; and (c), the hole and slot type resonator. The forms of these resonators, especially the parallel strip form of Fig. 21 (a), suggest that the resonators may be considered as sections of terminated transmission lines.

Voltage and current waves traveling down a section of uniform transmission line, terminated at one end by a short circuit and driven by a sinusoidal voltage at the other end, are reflected at the shorted end. The interference of the incident and reflected waves results in standing waves of voltage and current along the line. Since the voltage and current waves suffer phase changes on reflection differing by π radians, the corresponding standing waves are shifted by $\pi/2$ radians relative to one another. Thus the input admittance of the section of line is a periodic function of the distance, ℓ, to the shorted end. For a lossless line, this admittance is given by the expression:

$$Y = -jY_c \cot \frac{2\pi\ell}{\lambda} = -jY_0 \cot \frac{2\pi f \ell}{c}. \tag{26}$$

In (26), Y_o is the characteristic admittance of the line.   For a line of given length, this expression gives the input admittance as a function of frequency. When the frequency is $c/4\ell$, the line is a quarter wavelength long, and the
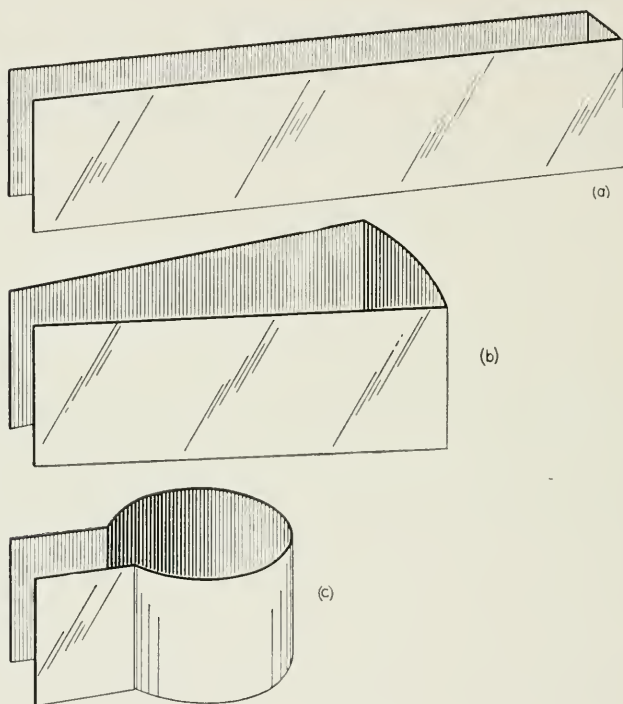
Fig. 21.—Three "strip type" cavities commonly used in magnetron resonator systems, each resonant at the frequency c/λ. Type (a) is essentially a quarter wavelength ($\lambda/4$) section of uniform transmission line.   It should be noted how types (b) and (c) are physically shorter than type (a) by virtue of the greater relative capacitive loading near the open ends.

input admittance, if the line is lossless, is zero.   In other words, the section of line resonates at this frequency.[19]   In a quarter wavelength resonator of

[19] The frequency of resonance, $f = c/4\ell$, is the fundamental or lowest frequency of an infinite series of resonant frequencies for which the line length is $(2q - 1)\lambda/4$;
$$f_q = (2q - 1)c/4\ell, q = 1, 2, 3, \cdots.$$
These frequencies may be specified by considerations of the phase relationships which must

uniform geometry, the voltage is maximum at the input end and the current is maximum at the shorted end, each varying sinusoidally to a node at the other end of the line.

The frequency of resonance of a section of terminated uniform line is thus determined by its length. If the geometry of the line is nonuniform, the frequency of resonance may be determined by the solution of Maxwell's equations with the appropriate boundary conditions. In general, this procedure is involved and tedious, however. One may get a reasonable idea of the values of ω_0 and Y_0 by assuming the half of the resonator near the open end to be capacitive only, the half near the closed end inductive only, and calculating C and L by application of elementary formulas to an equivalent parallel plate capacitance and a single turn sheet inductance of height h and proper cross sectional area. In the case of geometry like that of Fig. 21 (c), the division of the resonator on this basis is obvious.

A line of physical length ℓ', less than $\lambda/4$, may be made to resonate at the frequency c/λ by connecting across its input end a lumped capacitive susceptance, of magnitude, ωC, equal to that of the inductive susceptance of the line, $Y_0 \cot \dfrac{2\pi \ell'}{\lambda}$ [see equations (18) and (26)]. In like manner, resonators like those of Fig. 21 (b) and (c), whose physical length is less than $\lambda/4$, may be considered to be made up of an inductive section of uniform line across which additional capacitance has been inserted near the open end, bringing the frequency of resonance to c/λ. In Fig. 21, the three resonators of different physical lengths all resonate at the same frequency.

In addition to the resonant frequency of a resonator of distributed constants, one may define its Qs and characteristic admittance and link these to the rate of change of susceptance and energy storage capacity at resonance as was done for the circuit of lumped constants. Resonators of different geometry but of the same resonant frequency differ in characteristic admittance and loss conductance and hence in the Qs and the amount of energy which can be stored with unit voltage impressed across the input end. Of the resonator types shown in Fig. 21, the slot type has the largest admittance, the vane type, the smallest admittance, with the hole and slot type intermediate.

6. Resonator Systems

6.1 *Two Coupled Resonators:* The resonator system of the magnetron oscillator consists of a number of individual resonators of distributed parameters, machined into the anode block. As the simplest case of a system of coupled resonators, consider that having two resonators which are coupled

---

exist between oppositely traveling waves on a lossless line for constructive interference. These considerations are similar to those employed later in the discussion of the modes of oscillation of the magnetron resonator system as a whole.

only by the mutual linkage of magnetic lines and which resonate at the same frequency, ω_0, when uncoupled.

When such a coupled system is shock excited it is observed that the oscillation amplitude in either of the circuits is modulated at a so-called beat frequency, ω_B. A fraction or all of the energy in the system, depending on the initial conditions, surges back and forth at this frequency between the circuits similar to the manner in which the energy of motion is exchanged between two coupled pendulums. The total energy in the system is constant, the beats differing in phase by $\pi/2$ radians between the circuits.

The observation of beats is a manifestation of the fact that the two coupled resonators form a complex system oscillating simultaneously in its two modes for which the frequencies are $(\omega_0 + \omega_B)$ and $(\omega_0 - \omega_B)$. The oscillation in either circuit results from the superposition of the two component oscillations in this manner:

$$A \cos (\omega_0 + \omega_B) t + B \cos [(\omega_0 - \omega_B) t + \delta] =$$

$$(A - B) \cos (\omega_0 + \omega_B) t + 2B \cos \left(\omega_B t - \frac{\delta}{2} \right) \cos \left(\omega_0 t + \frac{\delta}{2} \right), \quad (27)$$

with a similar expression for the case when $A < B$. The oscillation may be predominantly of one frequency, that is, almost entirely in one mode, if, for example, $A \gg B$. In general, the oscillation is a superposition of a steady oscillation in the predominant mode $[(\omega_0 + \omega_B)$ if $A > B]$ and an oscillation whose amplitude varies with the beat frequency, ω_B. In the special case, when the component oscillations are of equal intensity, $A = B$, the amplitude of the resultant oscillation in either circuit goes to *zero* periodically at the frequency ω_B. This represents the case for which all the energy present in the system is transferred back and forth between the circuits.

The frequency separation of the two modes arises from the coupled effect of the oscillation in each of the circuits on the oscillation in the other. Thus, in the mode of lower frequency, $(\omega_0 - \omega_B)$, the two circuits oscillate in phase and the self induction effect in each circuit is aided by the mutual induction, each circuit behaving as though it were oscillating freely with a greater value of self inductance and hence at lower frequency [equation (17)]. For the mode of higher frequency, $(\omega_0 + \omega_B)$, the reverse is true. Here the two circuits oscillate out of phase by π radians, the mutual induction opposing the self induction and the circuits oscillating as though uncoupled with a smaller value of self inductance and hence at higher frequency.

If instead of shock exciting the system of two coupled resonators it is forced to oscillate by applying to it a sinusoidal voltage of variable frequency, the admittance of the system is found to pass through minima at the mode frequencies $(\omega_0 + \omega_B)$ and $(\omega_0 - \omega_B)$. Thus it is possible to drive the system and store energy in either of the two modes. For each mode of

the coupled circuit system, as for the simple single frequency resonator, Q parameters and a characteristic admittance may be defined which specify sharpness of resonance and energy storage capacity, respectively.

6.2 *The Multicavity Anode Structure:* As an introduction to the discussion of the multicavity resonator system of the magnetron oscillator having cylindrical symmetry, consider the system of a series of resonators machined side by side in a linear block as shown in Fig. 22.   One may consider such a linear array either to be infinite in extent or to be terminated in some manner at the ends of a string of N identical resonators.
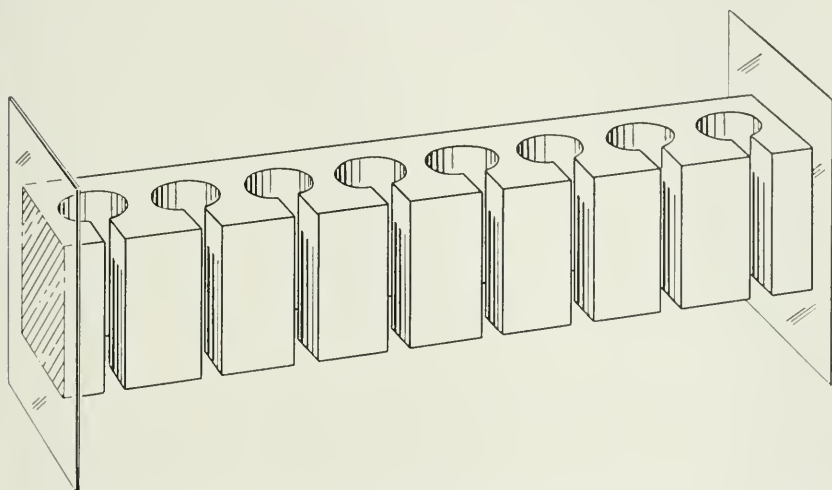
Fig. 22.—A linear array of resonators terminated at both ends by generalized terminations represented by planes.   The figure is meant also to indicate the nature of the infinite array of resonators referred to in the text.

The oscillation in each resonator of the array of coupled resonators is specified by a differential equation in terms of a variable, such as current or voltage, the constants of the circuit itself, and the mutual interaction between the circuit and its neighbors.   Each solution of the set of simultaneous differential equations for all the resonators involved corresponds to a definite phase shift between adjacent resonators.   The allowed values of this phase shift depend upon the boundary conditions imposed on the string of resonators.   If the block is infinitely long, all values of phase shift are allowed.   In terms of the electromagnetic field pattern formed on the front surface of the block by the fringing fields of the individual resonators, this means that traveling wave solutions representing waves of any velocity, traveling over the surface of the block in directions normal to the slots, are possible.   If the block is terminated, on the other hand, the boundary

conditions restrict the phase shift between resonators to a set of specific values. These correspond to the traveling waves which on reflection at the terminations constructively interfere.

The cylindrical magnetron anode structure is a series of N resonators connected in a ring. It may be thought to be a section of a linear array of resonators rolled into a cylinder. The boundary condition imposed is that of connecting together the resonators at the ends of the string. Under these circumstances only those modes of oscillation are possible for which the total phase shift around the ring is $2\pi n$ radians, n being any integer including zero. The oscillations in adjacent cavities then differ in phase by $\dfrac{2\pi n}{N}$ radians. Again this means that only those waves traveling around the anode block which constructively interfere are possible solutions. These are waves which, after leaving an assumed starting point and traversing the anode once, arrive back in phase with the wave then leaving in the same direction. The anode potential waves and the RF interaction fields in the interaction space to which they correspond have already been discussed in connection with equations (12) and (13). In these electromagnetic field patterns, the electric and magnetic field components are displaced both in space and time phase by $\pi/2$ radians relative to one another, similar to the manner in which voltage and current on a terminated transmission line are related.

6.3   *The Modes of the Resonator System:* It has been seen that the modes of oscillation of a magnetron resonator system are characterized by definite values of the phase shift between adjacent resonators specified by $\dfrac{2\pi n}{N}$, in which the parameter n may assume only integral values including zero. Each such mode of oscillation has a frequency different from the frequency of any other mode and from the frequency of one of the N resonators oscillating freely and uncoupled from its neighbors. In the general case of N coupled resonators, as in the case of two coupled resonators previously discussed, the modes of oscillation have different frequencies because of the effect of the mutual coupling between the resonators. For $N = 2$, the oscillations in the two resonators are either in phase or π radians out of phase, the induction in one circuit by the other either directly adding to or subtracting from the self induction. In the case of the multiresonator system the mutual induction effect may bear phase relations to the self induction other than 0 and π radians. Thus not only the magnitude of the coupling but also this phase relationship determines the magnitude of the effect of the mutual induction and hence the amount of deviation of the

mode frequency from that of a single uncoupled resonator.   If the coupling between resonators were in some way gradually reduced, all mode frequencies would converge to the value for a single uncoupled resonator.

The complete anode potential wave for a mode specified by the parameter n has been given in equation (12).   Each of its traveling components is represented by a fundamental of periodicity n and a series of so-called Hartree harmonics of periodicities $k = n + pN$, $p = 0, \pm 1, \pm 2, \cdots$. Any sinusoidal component for which the number of complete cycles around the anode is greater than $\frac{N}{2}$ is thus a harmonic of the complete field pattern for one of the modes whose fundamental is of periodicity $n = 1, 2, \cdots, \frac{N}{2}$. Physically distinguishable modes of oscillation exist only for the values of n less than or equal to $\frac{N}{2}$, including zero.   However, this accounts for only $\frac{N}{2} + 1$ of the N modes of oscillation which one expects a system of N resonators to possess.   The reason for this is that in general the frequency of a mode specified by the parameter n (except for the values 0 and $\frac{N}{2}$) is a double root for a perfectly symmetrical anode structure.   The mode is thus a doublet and is said to be degenerate.   One would expect this on mathematical grounds from the fact that the general solution of expression (12) has four arbitrary constants, whereas a singlet solution of the system of second order differential equations specifying the oscillations should have no more than two.

The nature of the degeneracy of the modes of the resonator system is perhaps most clearly seen by investigating what happens when the symmetry of the system is destroyed by the presence of a disturbance or perturbation at one point, a coupling loop in one of the cavities for example. Such a disturbance provides the additional boundary condition needed to remove the degeneracy.

Consider the effect of the perturbation on the nth mode.   It represents an admittance shunted across the otherwise uniform closed ring of resonators. This shunt admittance may be represented by ϵY_0, ϵ being a complex number, taken to be small for simplicity, and Y_0 the characteristic admittance of the closed resonator system.   In such a system, a potential wave incident upon the disturbance ϵY_0, having an amplitude a at the disturbance, breaks up into a reflected wave and a transmitted wave which, if ϵ is small, have the amplitudes $-\frac{\epsilon}{2} a$ and $\left(1 - \frac{\epsilon}{2}\right) a$, respectively [see equation (30)

for the reflection coefficient on a transmission line].　In passing across the disturbance, a wave undergoes a small phase shift $\delta\phi(\omega)$.　Since the total phase shift around the entire system must remain $2\pi n$, one may write: $\phi(\omega) + \delta\phi(\omega) = 2\pi n$, in which $\phi(\omega)$ is the phase shift around the system not including the disturbance.　If there is no disturbance, $\omega = \omega_n$ (ω_n is used here for the ω_0 of the nth mode), $\delta\phi(\omega) = 0$, and $\phi(\omega_n) = 2\pi n$.　For waves incident upon ϵY_0 from either direction, the respective amplitudes a and b at ϵY_0 must satisfy the equations:

$$a = \left[-\frac{\epsilon}{2} b + \left(1 - \frac{\epsilon}{2} \right) a \right] e^{j\phi(\omega)}$$

$$b = \left[-\frac{\epsilon}{2} a + \left(1 - \frac{\epsilon}{2} \right) b \right] e^{j\phi(\omega)}.$$

Writing $2\pi n - \left| \dfrac{\partial\phi}{\partial\omega} \right|_{\omega_n} \cdot (\omega - \omega_n)$ for $\phi(\omega) = 2\pi n - \delta\phi(\omega)$ in these equations and keeping only first order terms, one obtains the following pair of homogeneous linear equations for a and b:

$$\left[j \left| \frac{\partial\phi}{\partial\omega} \right|_{\omega_n} \cdot (\omega - \omega_n) - \frac{\epsilon}{2} \right] a - \frac{\epsilon}{2} b = 0$$

$$-\frac{\epsilon}{2} a + \left[j \left| \frac{\partial\phi}{\partial\omega} \right|_{\omega_n} \cdot (\omega - \omega_n) - \frac{\epsilon}{2} \right] b = 0.$$

These equations have a solution if and only if the determinant of the coefficients vanishes, that is, if:

$$j \left| \frac{\partial\phi}{\partial\omega} \right|_{\omega_n} \cdot (\omega - \omega_n) \left[j \left| \frac{\partial\phi}{\partial\omega} \right|_{\omega_n} \cdot (\omega - \omega_n) - \epsilon \right] = 0.$$

The two solutions are thus: $\omega = \omega_n$, for which $a = -b$, and $\omega = \omega_n - \dfrac{j\epsilon}{\left| \dfrac{\partial\phi}{\partial\omega} \right|_{\omega_n}}$ for which $a = b$.

From these facts concerning the amplitudes, namely, that at the disturbance ϵY_0 the amplitudes of the two oppositely traveling waves must either be equal or equal and opposite in sign for all time, one may conclude that the waves are of equal amplitude and that the standing waves resulting for each frequency must have a node at the disturbance for the $\omega = \omega_n$ solution and an antinode there for the $\omega = \omega_n - \dfrac{j\epsilon}{\left| \dfrac{\partial\phi}{\partial\omega} \right|_{\omega_n}}$ solution.

What this means in terms of the general solution for the degenerate mode

of periodicity n of equation (13) may now be seen. Expression (13) may be rewritten in terms of θ measured from the disturbance ϵY_0 in this way:

$$
V_{RF} = \sum_k (A_k - B_k) \cos (\omega t - k\theta + \gamma)
$$

$$
+ \sum_k 2B_k \cos \left(\frac{\gamma - \delta}{2} \right) \cos \left[\omega t + \left(\frac{\gamma + \delta}{2} \right) \right] \cos k\theta \quad (28)
$$

$$
+ \sum_k 2B_k \sin \left(\frac{\gamma - \delta}{2} \right) \cos \left[\omega t + \left(\frac{\gamma + \delta}{2} \right) \right] \sin k\theta.
$$

When the degeneracy is removed, the first summation representing the traveling wave vanishes since the amplitudes A_k and B_k are equal for the reasons indicated. Furthermore, the second term involving cos $k\theta$ terms is then the component of the doublet whose frequency changes, and the third term involving sin $k\theta$ terms is the component of undeviated frequency. For the latter component the disturbance appears at a node and hence has no effect. The reverse holds for the cos $k\theta$ solution. For it, the frequency deviation depends on the magnitude of the disturbance through the quantity ϵ. The disturbance caused by the coupling loop in an actual magnetron resonator system is sometimes sufficient to split the components into distinguishable resonances.

Thus an unsymmetrical multicavity resonator system in general has two modes of different frequency for each value of n. With respect to the asymmetry as origin, one of these modes has a cosine-like field pattern, the other a sine-like field pattern. This is true for $n = 1, 2, \cdots, \frac{N}{2} - 1$, contributing $N - 2$ modes. The remaining two modes of the resonator system, for which $n = 0$ and $\frac{N}{2}$, are singlet modes even in the symmetrical anode. This may be seen from the analysis demonstrating the splitting of the nth mode into two components. For the $n = 0$ mode, since the anode potential wave is independent of azimuthal angle, the solution $\omega = \omega_n$ for which $a = -b$ represents the trivial case of zero amplitude at all points. Similarly, for the $n = \frac{N}{2}$ mode (the π mode) the $\omega = \omega_n - \dfrac{j\epsilon}{\left| \dfrac{\partial \phi}{\partial \omega} \right|_{\omega_n}}$ solution

for which $a = b$ yields a cosine-like pattern giving zero potentials at each anode segment, an equally trivial case. Thus each of the N modes of the multicavity resonator system have been accounted for.

As an example, plots of the field configurations for the modes of a magnetron having eight resonators are shown in Fig. 23. For clarity, only the
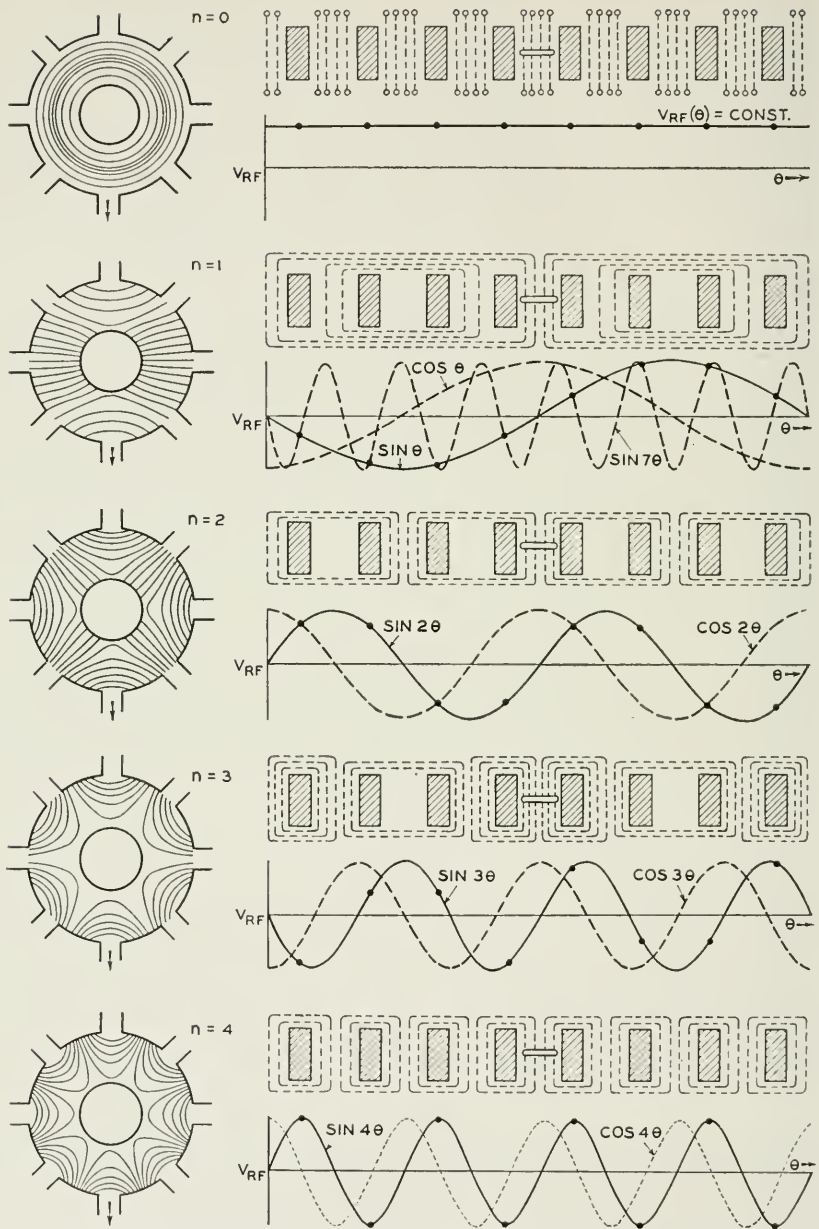
Fig. 23.—Configurations of electric fields, magnetic fields, and anode potentials for the $n = 0, 1, 2, 3$, and 4 modes of a resonator system having eight resonators. For each field pattern of periodicity n, the configuration of the electric lines of force in the magnetron interaction space is shown at the left, the configuration of the magnetic lines threading the resonators is shown at the upper right, and anode potential waves are shown at the lower

electric field lines of the fundamental component ($p = 0$) of each mode are shown in the interaction space.   Only the magnetic field lines are shown in the resonators.   Below these is plotted the distribution in potential for each of the fundamentals, sin $n\theta$ and cos $n\theta$, $n = 0, 1, 2, 3$, and 4.   For the $n = 0$ mode the magnetic flux threads through all the resonators in the same direction and returns through the interaction space.   That all the segments are in phase and the interaction space field is independent of angle may be seen.   That there is but one π mode is also seen from the fact that the cos 4θ term corresponds to zero potential on all the anode segments. The first Hartree harmonic for the $n = 1$ mode, namely that for which $p = 1$, having seven repeats ($k = 7$) or a total phase shift of 14π radians around the anode, is also plotted in Fig. 23 in addition to the fundamental. The fact that it yields the same variation of anode segment potential around the anode as the fundamental is apparent.

If the system of N resonators were shock excited it would oscillate in all of its modes simultaneously, producing beats in a manner analogous to but considerably more complicated than that for the system of two resonators already discussed.   Furthermore, if the system were forced to oscillate by an external drive whose frequency can be varied, the admittance of the system would go through a minimum at each of the mode frequencies.   With each such resonance there are associated values of the Qs, characteristic admittances, energy storage capacities, and the like.

The loading of the two modes of the same value of n by the output circuit of the magnetron depends on the position of the output loop relative to the respective standing wave patterns.   If the output coupling loop forms

---

right.   The interaction field plots represent only the fundamental components in each case.   The higher harmonics would affect the fields as plotted most radically near the slots in the anode surface.   The arrow shown in one of the slots in each case indicates the resonator which is coupled to the output circuit.   The field lines in each plot are spaced correctly relative to one another but not relative to those in any other plot.   In the plot of magnetic field lines in the resonator system (shown as dashed lines), the anode is developed from the cylindrical case, the anode segments being represented by the shaded rectangles.   At the center is a representation of the output loop.   The magnetic lines for the $n = 0$ mode thread through each resonator in the same direction and back through the interaction space in the opposite direction as indicated by the open circles at the ends of the lines.   For each mode the magnetic lines are shown for the instant when RF current flow is maximum and all anode segments are at zero potential.   In the plots of anode potential, the full lines represent the potential variations with azimuthal angle θ of the fundamental components sin $k\theta$, $k = n$. θ is measured from the position of the output coupling loop.   The full circles on these curves indicate the potentials of the anode segments.   The dashed lines represent the cos $k\theta$, $k = n$, modes.   It should be noted that the cos 4θ configuration is trivial as it yields zero potential on each anode segment at all times.   The cosine curves may also be taken to represent the azimuthal variation of magnetic field intensity which is in time quadrature with respect to the corresponding sine curves of potential.   Similarly, the sine curves may represent magnetic field intensity corresponding to the cosine curves of potential.   For the $n = 1$ mode the potential variation for the second Hartree harmonic ($n = 1$, $p = -1$) is also plotted (actually sin 7θ is plotted instead of sin -7θ for comparison with sin θ).   It is to be noted that it corresponds to the same anode segment potentials as its fundamental.

the asymmetry discussed above, one of the modes is strongly coupled and the other weakly coupled. The strongly coupled mode is that for which the anode potential wave is sine-like with respect to the cavity containing the output coupling loop as origin. In this mode the current and hence the magnetic flux in the output cavity is maximum. The fact that one of the modes of periodicity n is weakly coupled may be of significance in magnetron operation. As will be discussed later, oscillation in a loosely coupled mode, having a high Q by virtue of its being damped only by losses in the resonator system itself, may build up more rapidly under electron drive than that of the π mode. Then it is possible for the magnetron to oscillate either steadily under certain conditions, or intermittently, in an unwanted mode. For this reason it is usually necessary to provide a second asymmetry in the anode structure so as to shift the standing wave patterns of the two modes of same n most likely to offend and in this way to equalize their coupling to the output circuit.

6.4 *Higher Order Modes:* To this point in the discussion the RF circuit of the magnetron has been assumed to behave like a string of N lumped circuits coupled together in a ring, each circuit having only one natural frequency of resonance. This structure, as has been seen, has N modes of oscillation. Actually the multicavity resonator, since its constants are distributed, has an infinite number of modes of oscillation. They are to be distinguished by the nature of the variation of the RF field along the axis of the resonator system and radially in the individual resonators. Thus there may be nodal planes passing through the resonator system normal to its axis, or nodal cylinders, concentric with the axis of the system, passing through the resonators. The modes may be classified as symmetric or antisymmetric depending on whether the two ends of the system are in phase or π radians out of phase. The variation of RF voltage along the anode length is a circular sine or cosine function if the mode frequency is greater than the resonant frequency of the unstrapped resonator system and is a hyperbolic sine or cosine function if the mode frequency is less. The fundamental multiplet of N modes discussed above are symmetric modes corresponding either to no variation or to a hyperbolic cosine variation of RF voltage along the length. In these modes of the resonator system the cavities, considered as radial shorted transmission lines, resonate in their fundamental modes. Generally the frequencies of the higher order modes of the resonator system are quite far removed from those of the fundamental multiplet and only rarely need be considered.

6.5 *Other Types of Resonators:* As alluded to earlier, other types of magnetron resonators have been devised which can supply the proper alternate π mode potentials to the segments of a multisection anode. Two of these which have received some consideration by magnetron designers

but which have not come into general use will be mentioned in passing. One, the so-called "serpentine" anode structure, consists of a single slot, cut into the anode body, which winds up and down the anode length and around the interaction space. It is essentially a "half-section" wave guide, closed on itself, oscillating in its fundamental at the cut-off wavelength. As one passes along the resonator, the field for this mode is uniform, but, by virtue of the geometrical arrangement, the field it supplies to the interaction space is π radians out of phase from gap to gap. Other modes correspond to integral numbers of wavelengths along the length of the "serpentine" resonator. The separation in frequency between the fundamental and the next highest harmonic generally is not as great as is desirable.

The other magnetron resonator system to be mentioned involves the use of a single toroidal cavity of rectangular cross section whose inner cylindrical surface has been removed. Across this opening are placed the anode segments, adjacent ones being connected to opposite sides. The fundamental of this cavity corresponds to the cut-off wavelength as in the "serpentine" structure. This cavity has been mentioned in the literature[20] and has received some attention during the war. It is most useful in low voltage CW magnetrons where the small interaction space makes possible a resonator with sufficiently great mode separation between the fundamental and the first harmonic.

7. Separation of Mode Frequencies

7.1 *Necessity and Means:* The frequencies of the modes of the magnetron resonator system near that of the π mode would ordinarily be closely grouped were not steps taken to separate them. Curve (a) of Fig. 25 shows the distribution of mode wavelengths for a typical 10 cm. resonator system like that of Fig. 1. It is not easy to account for the exact nature of this curve. By virtue of the fact that the mutual induction effects between circuits bear different phase relations to the self induction effects for different modes one would expect the mode frequencies to differ. However, the conditions at the ends of the resonator block, in which region the flux lines link the resonators, are extremely important and affect frequency in a way not completely explained as yet. That the end regions should have an effect is understandable since they contribute capacitance and inductance as does any other part of the resonating cavity in which there are electric and magnetic field lines. Furthermore, as might be expected, it has been demonstrated that the end conditions have a greater effect on mode frequency the lower the mode number n. This results from the fact that the field strength in a mode of periodicity n falls off inversely as rapidly as the nth power of the distance from the resonator block.

From the point of view of the electronics of the magnetron, one might

[20] This resonator is the so-called "turbator" discussed by F. Lüdi, Bull. Schweiz. Elektrotechn. Verein, Vol. 33, No. 23 (1942).

think proximity of mode frequencies to be no problem because the different modes, even if of the same frequency, generally require different conditions relating the operating parameters V and B for oscillation [see relation (14)]. From the circuit point of view, however, close proximity of the mode frequencies is clearly undesirable, for under such conditions it is possible that a second mode may be excited by the electronically driven mode which is usually the π mode. The π mode oscillation, coupled to the second mode through some asymmetry in the resonator system, under these conditions sets up forced oscillations in the second mode. The interaction field pattern of the second mode then appears as a contamination of the π mode pattern, adversely affecting the electronic interaction with the π mode.

The mode frequency separation in magnetron resonator systems generally has been accomplished by two methods which bear an instructive relation to the means by which the mode frequencies of two coupled resonators may be separated. In the two resonator system, it is clear that the mode frequency separation can be increased by increasing the coupling since the difference in mode frequencies is brought about by the mutual coupling between the circuits, as has been explained. On the other hand, mode frequency separation may also be accomplished by detuning the individual resonators relative to one another. In either case the mode frequencies separate. Under shock excitation of the system, the beat frequency, equal to half the difference in the two mode frequencies, increases, corresponding to the greater rate of interchange of energy between one resonator and the other.

In the multicavity resonator system of the magnetron these means correspond, on the one hand, to the increase of coupling by conductive connections between the resonators, or so-called straps, and, on the other hand, to the use of cavities tuned alternately to different frequencies. This latter method has been used in the so-called "rising sun" anode structure to be discussed presently.

7.2 *Strapping of the Resonator System:* The idea of strapping a magnetron anode appeared in a British attempt to lock the oscillation of the resonator system into the π mode by connecting alternate anode segments together with wire straps. Although the number of modes of such a strapped structure is not changed, since its N-fold symmetry remains, the so-called "mode-locking straps" did succeed in separating the modes and making for easier oscillation in the π mode alone. The frequency separation of the modes is not infinite, however, because the straps are not of negligible length compared to a wavelength and thus have appreciable impedance between points on the structure to which they are connected. In most magnetron resonator systems today, straps of some form or other are employed.

In Fig. 24 are shown four types of strapping including the early British

SEC. A–A
SHOWING LOCATION
OF STRAPS IN TOP
OF ANODE
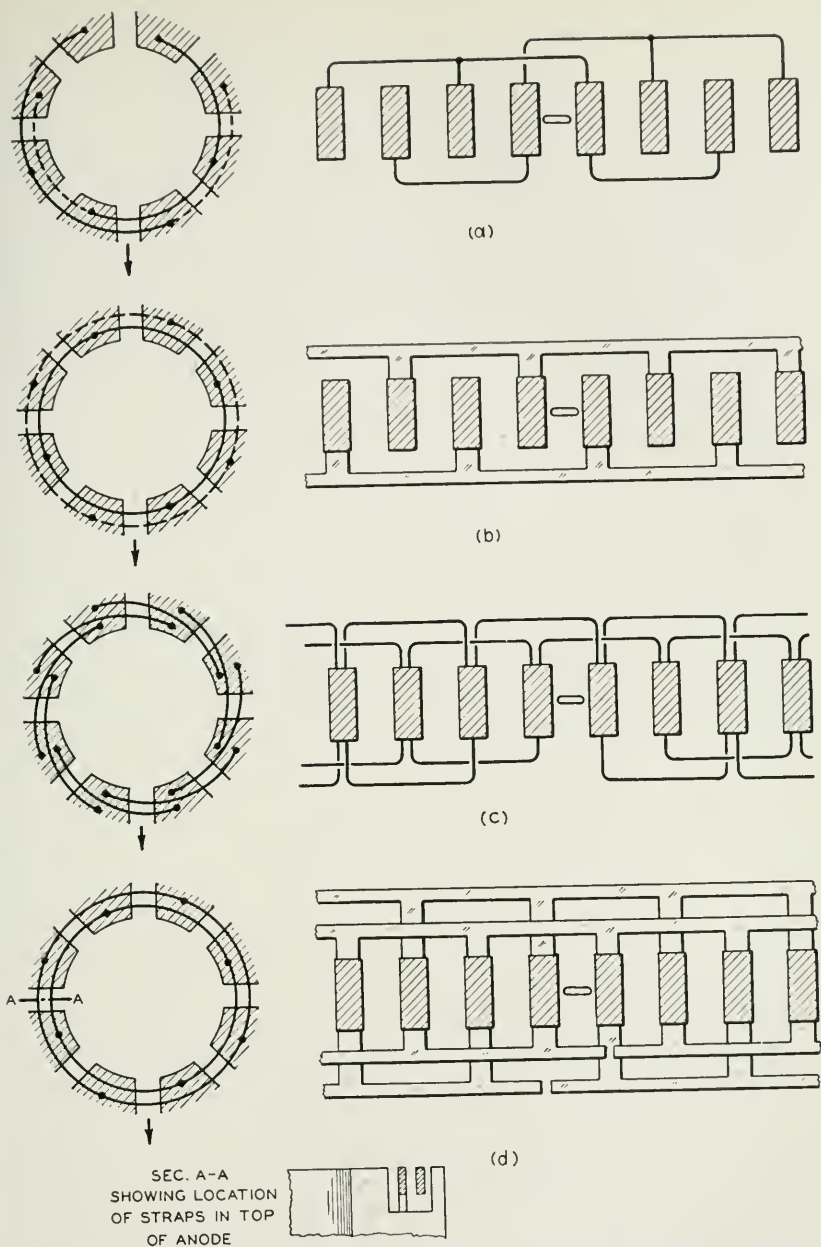
Fig. 24.—Schematic diagrams of four strapping schemes used in magnetron resonator systems. End views are shown on the left, and views rolled out as in Fig. 23 are shown on the right. (a) represents the early British type of unsymmetrical wire strapping, (b), a single ring type, (c), the so-called echelon type of wire strapping, and (d), the double ring type. In types (a) and (b) the straps on each end are at the same radius except where they overlap but are shown separated on the left-hand diagram for clarity. In types (c) and (d) the straps on the same end are at the same height and are shown as in the right-hand diagrams for clarity. The section A–A through type (d), shown below, indicates how the straps may be recessed into the resonator structure for shielding. The nature of the strap breaks introduced into types (c) and (d) are seen. In type (c) two links are removed and in type (d) actual breaks are made in the otherwise symmetrical rings. The breaks, shown here adjacent to the output coupling loop, are usually placed diametrically opposite.

223

type.　In Fig. 25 are shown the distributions of mode frequency for a typical resonator system unstrapped, and strapped with three of the types of strapping shown in Fig. 24.

It is possible to account, in quite simple terms, for the shift which takes place in the mode frequency distribution when the anode is strapped.　For this purpose consider a double ring strapped system like that of Fig. 24 (d). The role of the straps in determining mode frequency depends upon the relative magnitudes of their shunt inductive and capacitive effects.　The
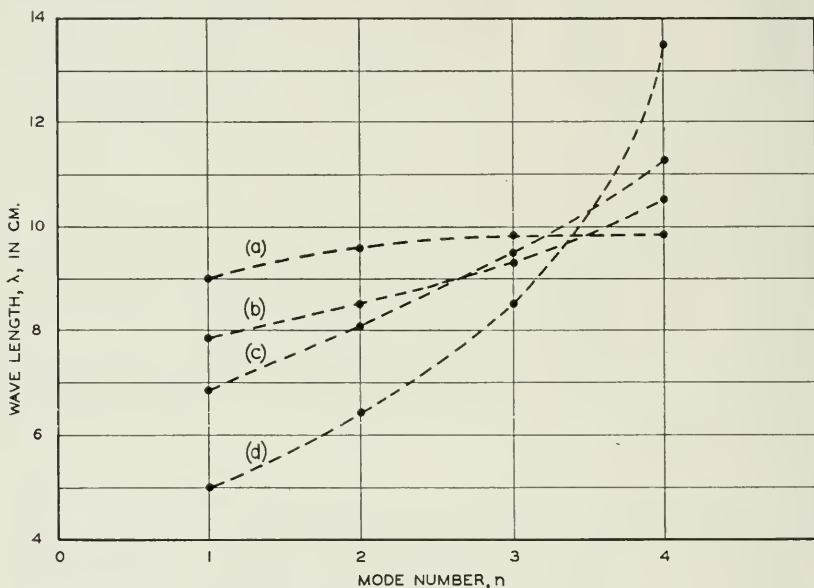
Fig. 25.—Plots of the variation of mode wavelength with mode number for a resonator system, unstrapped, or strapped in different ways.　Curve (a) is for the unstrapped anode structure.　Curve (b) is for the same structure strapped as shown in Fig. 24 (a), curve (c) for the same structure strapped as shown in Fig. 24 (b), and curve (d) for the same structure strapped as shown in Fig. 24 (d).　It is to be noted how the wavelength increases for large n and decreases for small n as the "strength" of strapping is increased.

capacitive effect of the straps for any mode depends upon the amount of shunt capacitance added relative to that already present in the resonators and upon the positions in the system to which they are connected.　This latter determines the average phase difference between the rings and thus their potential difference per unit RF voltage excitation.　Similarly, the inductive effect of the straps depends upon the magnitude of their shunting inductance relative to that in the resonators.　However, the important consideration concerning the points to which the straps are connected is the phase differences between points along the resonator system to which a

given ring is connected.   This determines the amount of current which the strap carries.

In the case of the π mode the two straps are π radians out of phase, each strap being connected to points which are in phase and at potential maxima [compare Figs. 23 and 24 (d)].   Their effect is predominantly capacitive. The only currents flowing in the straps are the charging currents of the strap capacitances.   If a resonator system having a total capacitance C, a total inductance L, and a π mode angular frequency ω_0, is strapped by a strapping system which adds a total capacitance C_s to the resonator system, the new frequency is

$$\omega_0' = 1/\sqrt{L(C + C_s)} = \omega_0/\sqrt{1 + C_s/C}.$$

The change in frequency is thus specified by the so-called "strength" or "tightness" of the strapping implied in the ratio of strap to resonator capacitance.

For modes of lower periodicity, $n < N/2$, the average potential difference between the straps and thus their capacitive effect is less because the straps connect points on the resonator structure differing in phase by less than π radians.   This corresponds to the shunting of a resonant line by a capacitance nearer the voltage node, at which point it would have no effect.   On the other hand, a *given* ring now connects points on the anode whose potentials differ in phase.   The ring thus provides additional conducting paths for the circulating RF currents in the resonator system.   These paths are essentially shunt inductances across the resonators which reduce the over-all inductance of the resonator system, shifting the mode to shorter wavelength.   As mode number decreases the two straps come closer together in potential but each strap connects points of greater potential difference.   Thus the capacitive effect decreases, the currents carried by the straps and hence the inductive effect increase, resulting in a progressive depression of mode wavelength.

In Fig. 25 the curves (a), (b), (c), and (d) represent the progression from the unstrapped case, (a), through three successive cases of increasing strength of strapping.   This increase in strength of strapping has resulted both in an increase of strap capacitance and a decrease of strap inductance as the increase of π mode wavelength and the decrease of mode wavelengths for smaller n demonstrates.   It is accomplished by increasing both the inter-strap and strap-to-body capacitances as well as the cross sectional area of the straps.

The magnitude of the inductance of a strap depends on its physical length between the points on the anode structure to which it is connected.   As this length increases, the strap inductance increases and hence has less effect as a shunt path.   For this reason, the effectiveness of a given strapping scheme in producing separation of mode frequencies is reduced if the anode

diameter is increased for higher voltage operation or to accommodate a greater number of resonators around the anode periphery.

Finally, it should be pointed out that the location of the straps at the ends of the anode structure causes their effectiveness to reduce with increasing anode length. As anode length is increased, the mode frequency distribution approaches that of the unstrapped resonator system of infinite length.

7.3 *Asymmetries in Strapping—Strap Breaks:* Of very great importance to the operation of a strapped resonator system is the degree of symmetry in the strapping system. The early British strapping is not symmetrical around the anode. The other three types shown in Fig. 24 are symmetrical except for breaks which are usually incorporated at least on one end of the anode. These asymmetries in the strapping provide the most convenient method of incorporating in the resonator system the additional asymmetry needed to orient the standing wave patterns of doublet modes with respect to the output circuit of the magnetron so as to equalize their loading. In addition, the strap asymmetries are arranged so as not to affect the symmetrical distributions of voltage and current in the resonator system for the π mode but to destroy such symmetry to an appreciable extent for other modes. For example, the destruction of the symmetry of the $n = 3$ mode pattern in a system of eight resonators by a single asymmetry in the strapping amounts to its contamination with a field component of periodicity corresponding to $n = 1$. Since the voltage and magnetic field values at which this contaminating component may be driven by the electrons are considerably farther removed from the π mode values than are those of the $n = 3$ mode, one has effectively converted the $n = 3$ mode into another mode less troublesome electronically.

In the echelon strapping of Fig. 24 (c) the asymmetry is produced by removal of two of the connecting links. The breaks in the strip or ring straps of Fig. 24 (d) are located as shown in the center of a link between two strap "feet". The break is thus at a current node in the strap for the π mode and has no effect on the pattern symmetry of this mode. For the other modes, however, the breaks appear at points where currents would normally flow and so represent sharp discontinuities in the structure for these modes.

In connection with strapped resonator systems there are two further points which should be mentioned. The first of these is the necessity in some cases of shielding the straps from the interaction space. The inner ring straps, mounted on the ends of the anode structure, present potentials to the interaction space which are independent of angle. This amounts to a perturbing $n = 0$ like component in the field pattern which is particularly strong near the ends of the anode and which affects magnetron operation adversely

in the manner already discussed in Section 4.5 *Effect of Other Components in the Interaction Field.* It can be removed by shielding the straps as shown for the case of the double ring straps in the section A-A of Fig. 24 (d). The shields, being parts of the anode segments, maintain the proper potential distribution in the interaction space. The echelon straps of Fig. 24 (c), since they overlap, need not be shielded.

Secondly, the location of the straps at the ends of the resonator system has an effect on the amount of variation of RF voltage along the length of the anode in the π mode. For anodes of length approaching a half wavelength or greater, this may become great enough to warrant attention. It is instructive, in this connection, to consider the anode structure as being made up of two circular strip transmission lines, the straps at the ends of the structure, between which there are connected at intervals N "half-section" wave guides, the N resonators of the system. Since the resonant frequency of the system is less than that of an individual resonator, the resonators act as inductances connected across each set of straps and as wave guide connections between the sets of straps operating beyond cut-off, the RF voltage varying hyperbolically along their lengths. Since the two ends of the anode structure are in phase, the mode is symmetric about the median plane and the variation of RF voltage along the axis is expressed by the hyperbolic cosine.

7.4 *The "Rising Sun" Resonator System:* The second type of magnetron resonator system in which the mode frequencies may be separated sufficiently well to allow "clean" operation in the π mode is an unstrapped structure involving the use of resonant cavities of two sizes so arranged that adjacent cavities are alternately large and small. This resonator system, called the "rising sun" system, accomplishes mode frequency separation by a means analogous to the increase in separation of the mode frequencies of a system of two coupled resonators achieved by relative detuning of the individual resonators. At the Columbia Radiation Laboratory during a series of experiments with asymmetries in an unstrapped resonator system, designed to achieve good operation in a harmonic of one of its doublet modes rather than in the π mode, it was observed that as the natural frequencies of the two sets of resonators were separated the mode frequencies diverged in two groups as though each group corresponded to one of the two sets of resonators and that in this configuration the π mode was quite well separated from the other modes. Thus the system appeared to be well suited for π mode operation, providing a means of mode separation without the use of straps. As such it is particularly adaptable to use in magnetrons of short wavelength where straps become very small and extremely difficult to construct.

A "rising sun" resonator system for $N = 18$ is shown in Fig. 26. The distribution of its mode frequencies is shown in Fig. 27, together with distributions for unstrapped and strapped eighteen resonator systems which have the same π mode frequency. It is seen that the modes of the "rising sun" resonator system arrange themselves into two groups or branches. Since the anodes are quite "long" (approximately $3\lambda/4$) the distribution for the unstrapped case having uniform resonator size [curve (a)] approximates the distribution for an anode of infinite length. It is reversed from the
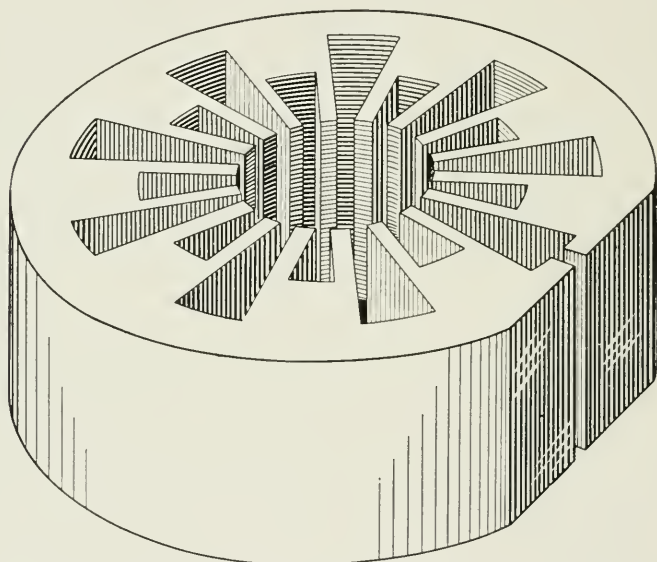
Fig. 26.—A so-called "rising sun" type resonator system having eighteen resonators. The slit at the "back" of the resonator in the right foreground is to be connected to the output circuit (compare Fig. 30).

unstrapped distribution of Fig. 25, curve (a), in which the end effects upon modes of small n are appreciable.

In the system of two coupled resonators, whether or not they are tuned individually to the same frequency, it has been seen that there are two mode frequencies corresponding to oscillations of the coupled system in which the resonators are in phase or π radians out of phase. In the "rising sun" system one observes two *groups* of mode frequencies corresponding to oscillations of the coupled *systems* of resonators in which corresponding

modes of the component systems, that is, modes of the same pattern periodicity, add in phase or π radians out of phase. Thus the modes of a "rising sun" system having eighteen resonators corresponding to $n = 0, 1, 2, \cdots,$ 8, and 9 are to be compounded of the two sets of modes for the large and
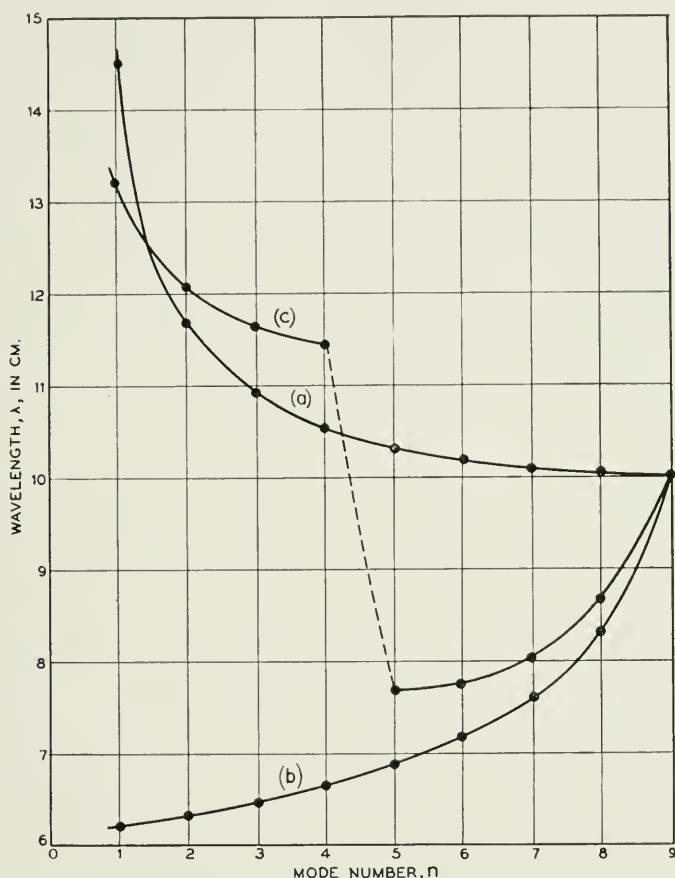
Fig. 27.—Plots of mode wavelength as a function of mode number for three different resonator systems of eighteen resonators having the same axial length and π mode wavelength. Curve (a) is for an unstrapped resonator system of identical resonators. Curve (b) is for a "heavily" strapped resonator system having identical resonators. Curve (c) is for a "rising sun" type resonator system having a ratio of resonator wavelengths of 1.8.

small resonators, each set including modes of periodicities $n' = 0, 1, 2, 3$, and 4. The two $n' = 0$ modes when added in phase give the $n = 0$ mode of the whole system but when added π radians out of phase give the $n = 9$ or π mode of the whole resonator system. This latter fact is perhaps made more clear by the observation that in the $n = 9$ or π mode all the large

resonators are in phase and π radians out of phase with all the small reso-
nators. Similarly the $n' = 1$ modes of the two component sets of reso-
nators added in phase give the $n = 1$ mode, added π radians out of phase
give the $n = 8$ mode. The $n' = 2$ modes yield the $n = 2$ and $n = 7$ modes
of the total resonator, and so on. Modes of the component resonator systems
of different periodicities do not add as they are uncoupled in the same sense
as two modes of a resonator system with resonators all of the same size.
The curve showing the distribution in mode frequency from $n = 0$ to $n = 4$
thus has the usual shape for increasing periodicity of the field pattern.
The distribution in mode frequency for the remaining modes, however, is
reversed in form and, as n goes from 9 down to 5, appears as a distribution
should for which the mode periodicity increases. The two branches of the
mode frequency distribution curve thus appear as approximate mirror
images which are shifted relative to one another along the frequency scale
by virtue of the difference in phase of the mutual coupling between the two
sets of resonators. As far as frequency is concerned, the π mode of the
total system has the characteristics of a mode whose field pattern is inde-
pendent of angle, and its frequency is well separated from those of other
modes.

As in the case of both unstrapped and strapped symmetrical resonator
systems, equivalent circuits have also been devised and studied for the
"rising sun" structure. Suffice it to say here concerning them that in each
case it has been possible to explain and predict the mode frequency behavior
to a surprising degree of accuracy.

As a magnetron resonator the "rising sun" system has both advantages and
disadvantages. Its most obvious advantages are its lack of strapping with
consequent ease of construction for short wavelengths and the ability to
make an anode structure of any length with no penalty in mode frequency
separation. Although the frequency separation of the π mode from other
modes is not as great as is possible in strapped magnetrons, its independence
of anode length and the fact that it can be realized at higher values of N are
both important for high power magnetrons. Furthermore, the "rising sun"
structure, having no strap losses, possesses an inherently higher unloaded Q
than strapped resonator systems. This results in an improvement in circuit
efficiency by a factor which may be as high as 1.2 at 1.25 cm. wavelength.

The major disadvantage of the "rising sun" resonator system is the
presence in its π mode interaction field of a strong admixture of a com-
ponent independent of angle. How this comes about may be seen from the
following considerations: The π mode frequency of the composite resonator
system lies somewhere between the free oscillation frequencies of the large
and small resonators. When oscillating in the π mode, therefore, the large
resonators are longer and the small resonators shorter than an equivalent

quarter wavelength. Said another way, the electrical distance from the
back of a large resonator to its opening in the anode, across the segment face,
and to the back of the adjacent small resonator is an electrical half wave-
length along which the voltage and current vary approximately sinusoidally.
For this reason the maximum in the RF voltage and the corresponding cur-
rent node do not appear at the mouth of either cavity but at a point (M of
Fig. 28) somewhat inside the opening of the large cavity. This means that
the electric field across the mouth of the larger cavity is greater than that
across the mouth of the smaller cavity. The excess, since all the large
cavities are in phase, adds up around the anode to form an electric field
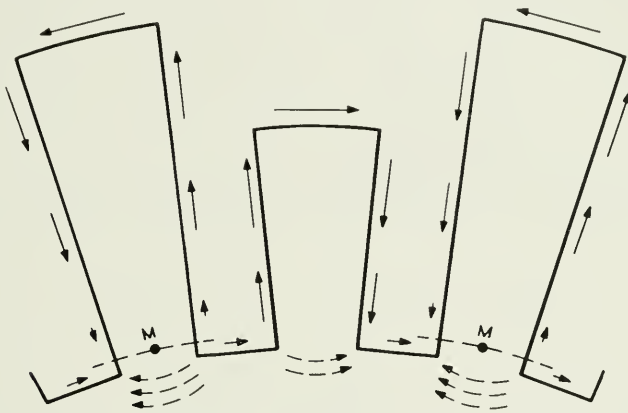
Fig. 28.—A diagram illustrating the origin of the component independent of angle in the
π mode interaction field of the "rising sun" resonator system. The length of the arrow
lying parallel to and just inside the resonator wall at any point is proportional to the magni-
tude of the RF current flowing there. The RF current node, and hence RF voltage maxi-
mum, occurs at M, inside the mouth of the larger cavity. Note that this gives rise to
currents flowing in the faces of the anode segments which are in the same direction around
the interaction space and that the RF field strength across the mouth of a larger cavity is
greater than that across the mouth of a smaller cavity. This latter is indicated schemati-
cally by means of the dashed arrows.

component independent of angle like the $n = 0$ mode field in Fig. 23.
Further, it is seen in Fig. 28 that at any instant there are currents flowing
across the faces of the anode segments which are all in the same direction
around the anode. With this net circumferential current is associated the
unidirectional RF magnetic field component parallel to the axis shown
also in Fig. 23.

The amount by which the two sets of mode frequencies of the "rising
sun" structure are separated increases with increasing ratio of resonator
sizes. Corresponding to this, the amount of $n = 0$ like component in the
interaction field increases.

The presence of the component independent of angle in the interaction

field is thus an inherent characteristic of the "rising sun" resonator system. One is faced with the problem of designing for sufficient mode separation without unduly increasing this component. How the presence of this component in the interaction field perturbs the electronic interaction with the π mode, resulting in a performance characteristic like that of Fig. 20, has already been discussed.

8. Output Circuit and Load

8.1 *Output Circuit:* In the general physical description of the centimeter wave magnetron whose constituent parts are shown in Fig. 1 there remains the discussion of the output circuit. The output circuit is the means of coupling the fields of the magnetron resonator to the load and as such it must contrive to induce a voltage across a coaxial line or a waveguide to which the load circuit is connected. Several types of coupling are involved in magnetron construction. These are illustrated schematically in Fig. 29. Here the resonator of the magnetron is represented by a simple $L - C$ circuit and any transformer action of the output circuit between the resonator and the load is to be accounted for by the unspecified network T. The scheme of Fig. 29 (a) involves magnetic coupling, that of (b), electrostatic coupling, those of (c) and (d), two forms of direct coupling.

Type (a), it is clear, corresponds to the output coupling accomplished by a loop, like that shown in Fig. 1, feeding a coaxial line. The loop may be placed inside the cavity as in Fig. 1, may be placed above the resonator in the end space as in the case of the so-called "halo" loop, or may be placed with its plane parallel to the axis of the anode between the resonators in the end space. In each case the coupling is effected mainly by linking of magnetic lines of force by the loop. The coupling is not entirely magnetic, however. There is electrostatic induction in the loop by the anode segments near it, corresponding to coupling of type (b) of Fig. 29, and in the case of the third possible placement of the loop listed above, there is involved some direct coupling of the type (c) of Fig. 29 since the loop is terminated on an anode segment on which there is RF potential.

In most cases the coaxial line must expand in dimensions from the loop extremity, pass through the vacuum envelope, and be provided with a means of coupling to the coaxial load line of the system in which the magnetron is used. The output circuit from the loop to the smooth line of the system must provide the transformer action necessary to load the loop by an admittance which gives the desired Q [see equation (23)]. What this admittance must be is dependent, to be sure, on the size and position of the loop, that is, upon the degree to which it couples the magnetic lines in the resonator. Generally, the attempt is made to build the transformer

into the magnetron, preferably inside the vacuum envelope where any large standing waves present are less likely to cause RF voltage breakdown.
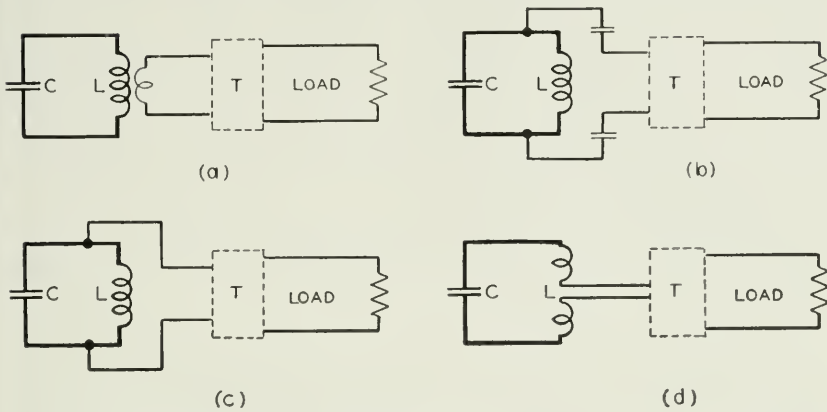
Fig. 29.—Schematic circuit diagrams representing four types of output couplings Type (*a*) is magnetic coupling, (*b*), electrostatic, (*c*) and (*d*), two forms of direct coupling. The unspecified network T represents the output circuit between the points where it couples to the resonator system and the load.
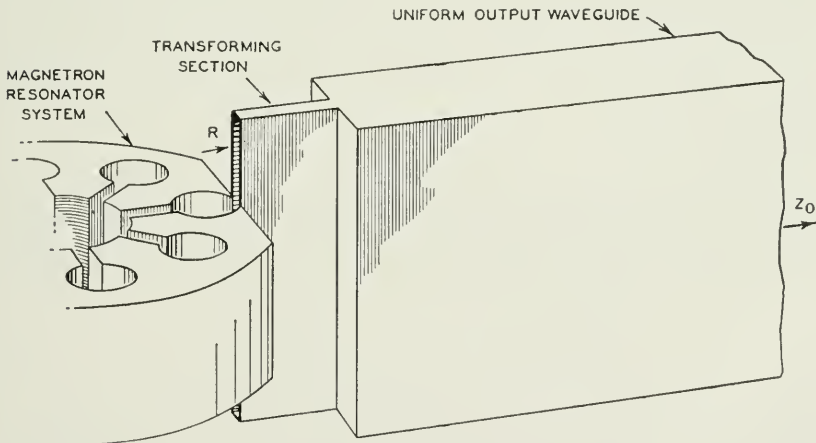
Fig. 30.—An example of a type of waveguide output circuit. It is representative of the type of coupling of Fig. 29 (*d*). Other types of resonator systems may be used (compare Fig. 26), and the transforming section, for example, may be of dumbbell-shaped cross section rather than of rectangular cross section as shown.

The type of magnetron output circuit represented schematically by Fig. 29 (d) is the so-called waveguide output. Here one of the resonators is broken into by means of a slit as shown in Fig. 30. Attached to this slit is a transforming section which feeds directly into the load waveguide. The im-

pedance, R, at the resonator presented by the output circuit must be small. The characteristic impedance, Z_0, of the waveguide is large. The transformer usually consists of a quarter wavelength section of characteristic impedance equal to the geometric mean of R and Z_0, $\sqrt{RZ_0}$. The vacuum seal is made by a dielectric window in the waveguide. The quarter wave transformer section may be a parallel plate transmission line cut in the resonator block or may be a waveguide line of rectangular or dumb-bell shaped cross section. Here again, the specific amount of its transformer action must be adjusted, usually by variation of the small dimension, until it provides the proper value of Q_{ext}.

Another type of output circuit involves coupling a coaxial line directly onto the straps. This represents practically a pure case of the type of coupling shown in Fig. 29 (c).

8.2   *Load:* The load admittance which the output line presents to the output circuit of the magnetron oscillator depends upon the characteristic admittance, Y_0, of the line and upon the manner in which the line is terminated. In discussing the single resonator of the magnetron resonator system as a section of lossless transmission line terminated by a short circuit, the input admittance and its relation to the standing waves on the line were mentioned. Since the termination reflects all the energy incident upon it in the shorted line, the voltage standing wave ratio, σ, defined as the ratio of the maximum voltage to the minimum voltage along the line, is infinite. The input admittance of a shorted section of length ℓ has been given in equation (26).

In the general case in which the line is terminated by an admittance, Y_T, not all the energy incident upon the termination is reflected, the standing wave, whose position is determined by the phase of Y_T, has a finite value of σ greater than unity, and the input admittance is given by the expression

$$Y = Y_0 \frac{Y_T + jY_0 \tan \dfrac{2\pi\ell}{\lambda}}{Y_0 + jY_T \tan \dfrac{2\pi\ell}{\lambda}}. \tag{29}$$

If the voltage reflection coefficient, $\vec{r}$, is defined as the ratio of the complex voltage amplitudes of the reflected and incident waves, A_R and A_I,

$$\vec{r} = \rho e^{j\phi} = \frac{A_R}{A_I},$$

the standing wave ratio may be written

$$\sigma = \frac{|A_I| + |A_R|}{|A_I| - |A_R|} = \frac{1 + \rho}{1 - \rho},$$

and the input admittance, Y, expressed as

$$Y = Y_0 \frac{1 - \vec{r}}{1 + \vec{r}}.$$

Conversely:

$$\vec{r} = \rho e^{j\phi} = \frac{\sigma - 1}{\sigma + 1} e^{j\phi} = \frac{1 - Y/Y_0}{1 + Y/Y_0}. \tag{30}$$

If the line is matched, that is, terminated in its characteristic admittance, $Y_T = Y_0$, it is clear that the input admittance, Y, is equal to Y_0, the voltage reflection coefficient, $\vec{r}$, is zero and the voltage standing wave ratio, σ, is unity.

These concepts are recalled here because they are used in specifying the magnetron load. The remaining point of interest with respect to admittance relationships on transmission lines is the transformation of admittance which occurs in going through a line of variable characteristics such as the output circuit of the magnetron. Such a section of nonuniform line may in general be considered as a lossless transducer, the admittance transformation through which is expressed as a bilinear form. In terms of the reflection coefficient $\vec{r}_2$ looking into the load at the output terminals of the transducer, the reflection coefficient $\vec{r}_1$, looking into the transducer at its input terminals may be written thus:

$$\vec{r}_1 = e^{-j\vec{\alpha}_{12}} \frac{\beta_{12} + \vec{r}_2 e^{j\vec{\alpha}_{21}}}{1 + \beta_{12} \vec{r}_2 e^{j\vec{\alpha}_{21}}}. \tag{31}$$

In this expression the number β_{12} and the angles $\vec{\alpha}_{12}$ and $\vec{\alpha}_{21}$ are the three parameters completely describing the transducer, which for lossless transducers are real numbers.

9. EQUIVALENT CIRCUIT THEORY

9.1 *The Equivalent Circuit:* From time to time in the discussion thus far, reference has been made to a lumped constant circuit or circuits which may be considered to be equivalent to the resonator system, output circuit, and load of the magnetron oscillator. One is now in a position to appreciate the justification for the use of such a simple, singly resonant circuit to represent as complex a device as the magnetron. As has been pointed out, this justification lies in the ability to separate the mode frequencies and to diminish sufficiently well, excitation of all modes except the π mode. Further, in the discussion of the equivalent circuit shown in Fig. 2, it was pointed out how the output circuit and the electronics may be treated by circuit analysis. This particular equivalent circuit will now

be discussed in some detail. From an analysis of this circuit it will be explained how the power which the magnetron delivers and the frequency at which it oscillates depend upon the load attached to it.

Consider now the equivalent circuit shown in Fig. 2, or as repeated in Fig. 31 (a). Since the N cavities of which the magnetron resonator system is composed are essentially in parallel for the π mode, the C of the equivalent circuit is N times the capacitance and the L is $\dfrac{1}{N}$ times the inductance of a single cavity. G_c is the shunt conductance representing the series resistance
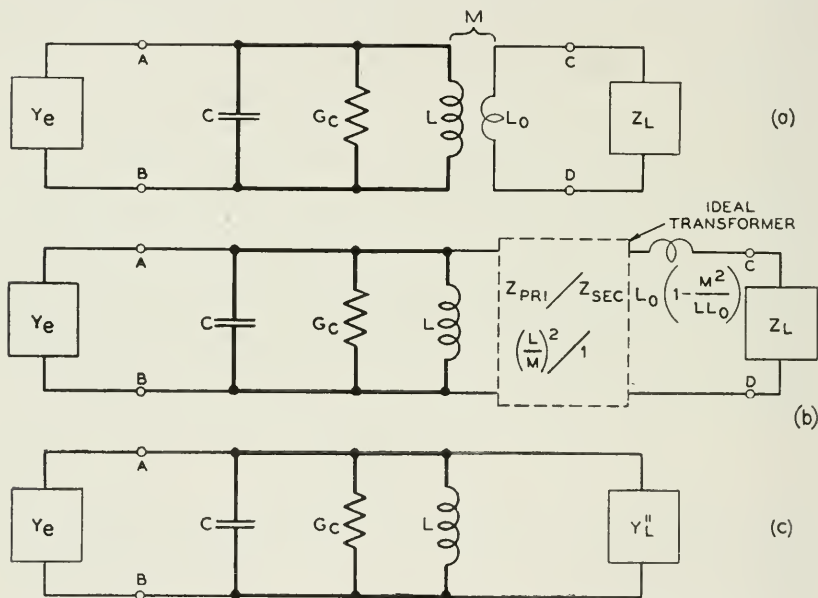
Fig. 31.—A diagram showing an equivalent RF circuit for the magnetron oscillator, (a), and how this circuit may be reduced in two steps, (b) and (c), to a simpler form.

in the copper walls of the resonator system, L_0, the inductance of the output loop which is coupled by the mutual inductance, M, to the lumped inductance of the equivalent resonating circuit. Z_L represents the impedance of the load at the loop terminals. Thus it represents the load impedance to which the magnetron is attached, transformed through the output circuit to the loop terminals.

The first step in understanding the circuit of Fig. 31 (a) is to reduce it to a simpler form. This is done in two steps as shown in Fig. 31. The inductances L and L_o, with mutual inductance M between them, form a transformer through which the impedances in the secondary circuit, $j\omega L_0$ and

Z_L, are reflected into the primary circuit. It may be shown[21] that the circuit of Fig. 31 (b) is the equivalent of that of Fig. 31 (a). The coupling into the primary circuit is represented by an ideal transformer connected across the primary inductance L to the secondary winding of which are connected the load impedance Z_L and a reduced loop reactance

$$X_0 = j\omega L_0 \left(1 - \frac{M^2}{LL_0}\right).$$

The ideal transformer effects a voltage transformation of $\frac{L}{M}$:1 or an admittance transformation of $\left(\frac{M}{L}\right)^2$:1 from its secondary to its primary terminals. Thus, the admittance, Y_L'', presented at the primary terminals of the ideal transformer, is

$$Y_L'' = G_L'' + jB_L'' = \left(\frac{M}{L}\right)^2 \left(\frac{1}{X_0 + Z_L}\right)$$

$$= \left(\frac{M}{L}\right)^2 Y_L' = \left(\frac{M}{L}\right)^2 (G_L' + jB_L') \tag{32}$$

in terms of the admittance Y_L' at the secondary terminals.

The equivalent circuit has now been reduced to that of Fig. 31 (c) used earlier in the discussion of a single resonator of the magnetron resonator system. Each of the quantities defined or derived for this circuit are now to be applied to the magnetron resonator system as a whole. These include the characteristic admittance of the resonator, Y_0, the unloaded, loaded, and external Qs given by the relations (21), (22) and (23), and the circuit efficiency, η_c, of equation (25).

Looking to the left at the terminals AB into the electron stream one sees the electronic admittance $Y_e = G_e + jB_e$. This is defined in terms of the current, I_{RF}, induced in the anode segments by the electrons moving in the interaction space, and the RF voltage, V_{RF}, appearing across the resonators, that is, across the terminals AB of Fig. 31 (c). Looking into the circuit at the terminals AB one sees the admittance Y_s.

This admittance by equations (19) and (32) is

$$Y_s = G_c + j2Y_{0c}\frac{\omega - \omega_0}{\omega_0} + Y_L''$$

$$= G_c + j2Y_{0c}\frac{\omega - \omega_0}{\omega_0} + \left(\frac{M}{L}\right)^2 (G_L' + jB_L'). \tag{33}$$

[21] See Guillemin, Communications Networks, Vol. II, p. 154.

The condition for oscillation stated earlier requires that:

$$Y_e + Y_s = 0,$$

or:

$$G_e + jB_e + G_e + j2Y_{0c}\frac{\omega - \omega_0}{\omega_0} + \left(\frac{M}{L}\right)^2(G'_L + jB'_L) = 0. \quad (34)$$

Separating real and imaginary parts, this reduces to the pair of equations:

$$\left.\begin{array}{l} G_e + G_s = 0, \quad\quad \text{or,} \quad\quad G_e + G_c + \left(\frac{M}{L}\right)^2 G'_L = 0, \\[4mm] B_e + B_s = 0, \quad\quad \text{or,} \quad\quad B_e + 2Y_{0c}\frac{\omega - \omega_0}{\omega_0} + \left(\frac{M}{L}\right)^2 B'_L = 0. \end{array}\right\} \quad (35)$$

Since $\left(\frac{M}{L}\right)^2 G'_L$ is equal to G''_L, one may write the first of equations (35) in terms of the Qs defined by (21), (22), (23) and (24) thus:

$$-G_e = G_s = Y_{0c}\left(\frac{1}{Q_0} + \frac{1}{Q_{ext}}\right) = \sqrt{\frac{C}{L}}\frac{1}{Q_L}. \quad (36)$$

9.2 *The Rieke Diagram:* The electronic conductance and susceptance, being functions of the parameters such as V_{RF}, V, and B which govern the electronic behavior of the magnetron, are not known *a priori* except for the fact that they are undoubtedly slowly varying functions of frequency. The circuit conductance and susceptance are given by equation (33). Equations (35) state that the circuit conductance and susceptance must be the negative of the electronic conductance and susceptance respectively. It is from these relations that the behavior of the oscillator under changes of load is to be inferred.

Suppose now that one were to vary the load in such a way that only B'_L changes, G'_L remaining constant. Then, by the first of equations (35), G_s and hence G_e would not change. If, further, the frequency of oscillation changes such that

$$2Y_{0c}\frac{\Delta\omega}{\omega_0} + \left(\frac{M}{L}\right)^2 \Delta B'_L = 0, \quad (37)$$

by equations (35) B_s and hence B_e remain constant as well and the electronic operation of the magnetron involving the RF voltage, V_{RF}, is undisturbed. Since the power delivered by the magnetron is

$$P = -G_e V^2_{RF} = G_s V^2_{RF}, \quad (38)$$

contours of constant G'_L on a plot of performance versus load, Fig. 32 (a), are thus contours of constant output power. Along any such constant

power contour the frequency of the magnetron varies linearly with B'_L as equation (37) indicates. Hence any constant frequency contour on the diagram is obtainable from a neighboring contour of different frequency through translation in the direction of B'_L by an amount given by equation (37). The form of the contours of constant frequency depends upon the interdependence of the electronic parameters G_e and B_e. The fundamental electronic performance as a function of load is specified by the conductance G_s presented to the electrons at the anode slots, changes in load susceptance being compensated for by frequency changes and hence susceptance changes in the resonator. As G'_L is varied G_s, G_e, and the output power must vary as well. If B_e is independent of these changes the constant frequency contours will correspond to lines of constant B'_L. Actually it is found that B_e does depend to some extent on G_e, resulting in the constant frequency contours on the $G'_L - B'_L$ plot being approximately straight lines inclined to the constant B'_L lines at a small angle α as shown in Fig. 32 (a).

If the $G'_L - B'_L$ plane is transformed to the reflection coefficient or $\vec{r}$ plane on which the load characteristic is usually plotted, contours of constant G'_L or power become circles tangent to the circle $\rho = 1$ at the same point. Constant B'_L contours form the set of circles orthogonal to these. The contour of constant frequency of Fig. 32 (a) transforms to a circle which intersects all the contours of constant power at the angle α.

Fig. 32 (b) is thus the form of the characteristic depicting the dependence of magnetron performance on load, called the Rieke diagram, plotted for that point in the equivalent circuit where the admittance looking out into the load is $Y'_L = G'_L + jB'_L$. Between this point and a point in the smooth output line, with respect to which the load admittance is usually measured, there is the series reactance X_o and the output circuit of the magnetron, forming a transducer through which the reflection coefficient defining the load admittance may be transformed by the expression (31). In undergoing such a transformation the contours of constant power and frequency plotted on the $\vec{r}$ plane retain their general form although they may be rotated and expanded or contracted. Thus the general form of the Rieke diagram shown in Fig. 32 (b) should be the same as that experimentally determined and plotted on a reflection coefficient plane for a point in the output line. Fig. 33 is such a Rieke diagram and its resemblance to that of Fig. 32 (b) is apparent.

9.3 *Magnetron Circuit Parameters:* The fact that the circuit theory of the magnetron based on the simple equivalent circuit of Fig. 31 (a) explains the nature of the Rieke diagram so well is ample justification for its use. The parameters which specify the equivalent magnetron circuit may be

$$\vec{r} = \rho e^{j\phi} = \dfrac{1 - \left[\dfrac{G'_L}{Y_0} + j\,\dfrac{B'_L}{Y_0}\right]}{1 + \left[\dfrac{G'_L}{Y_0} + j\,\dfrac{B'_L}{Y_0}\right]}$$

(a)

(b)
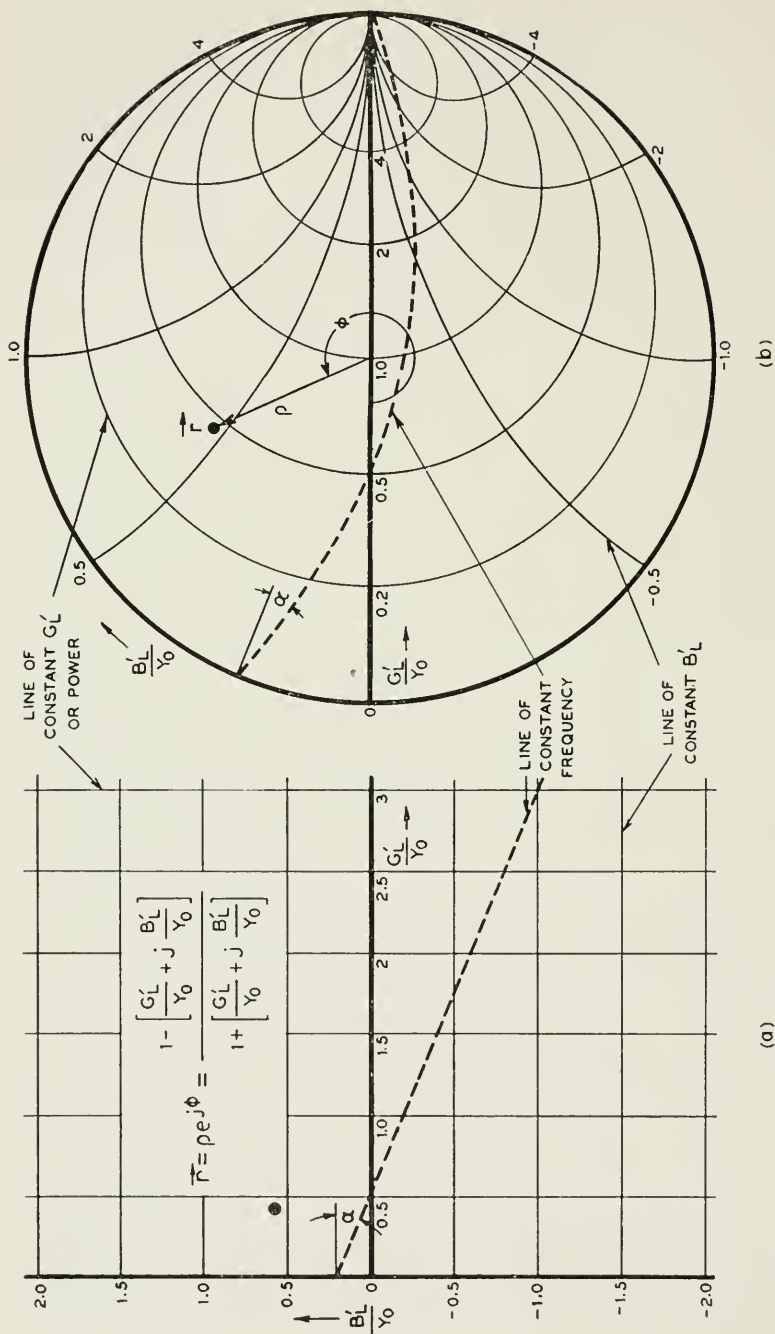
Fig. 32.—Charts showing how the G'_L-B'_L plane of (a) transforms to the $\vec{r}$ plane of (b) under the bilinear transformation written out in the upper half of chart (a). A line of constant magnetron frequency is shown plotted on chart (a) and in its transformed position on chart (b). Note that all the straight lines of (a) becomes circles in (b) with angles of intersection preserved. Note also the transformation of the admittance point represented by the filled circle. The configuration of constant G'_L lines, corresponding to constant output power, and constant f lines shown on chart (b) is to be compared with the experimental Rieke diagram of Fig. 33.

determined as a function of frequency by a measurement of the impedance Z_c looking into a non-oscillating magnetron through its output circuit and an independent measurement or calculation of one of the parameters such as C or Y_{0c}. The impedance Z_c in terms of the equivalent circuit is the impedance across the terminals CD in Fig. 31 (a) looking to the left with

CONTOURS OF CONSTANT POWER OUTPUT
CONTOURS OF CONSTANT FREQUENCY
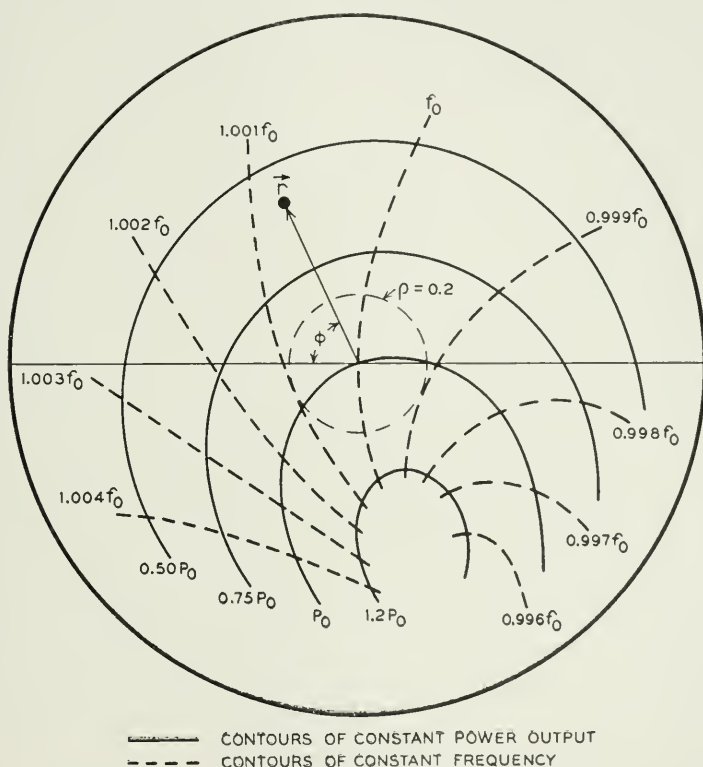
Fig. 33.—A typical experimental Rieke diagram for magnetrons in the centimeter wavelength range. The $\rho = 0.2$ circle indicates a pulling figure of $0.02\, f_o$ mc/s for this example. Present practice is to couple long wavelength magnetrons more tightly to the load and short wavelength magnetrons less tightly to the load than this. The rotation of the diagram from the position of the contours of chart (b) of Fig. 32 is attributable to the transformer action between the terminals of the ideal transformer of the circuit of Fig. 31 (b) and the point in the output line to which the experimental diagram is referred.

the terminals AB open circuited. From the circuit of Fig. 31 (b) this is seen to be

$$Z_c = j\omega L_0\left(1 - \frac{M^2}{LL_0}\right) + \left(\frac{M}{L}\right)^2\left[\frac{1}{G_c + j2Y_{0c}\dfrac{\omega - \omega_0}{\omega_0}}\right] \tag{39}$$

which, using equation (21), becomes

$$Z_c = jX_0 + \left(\frac{M}{L}\right)^2 \frac{1}{Y_{0c}} \left[\frac{1}{\dfrac{1}{Q_0} + j2\,\dfrac{\omega - \omega_0}{\omega_0}} \right] = R_c + jX_c \qquad (40)$$

The experimental data may be given in terms of the values of R_c and X_c plotted as functions of ω as shown in Fig. 34. The value of X_0 is deter-
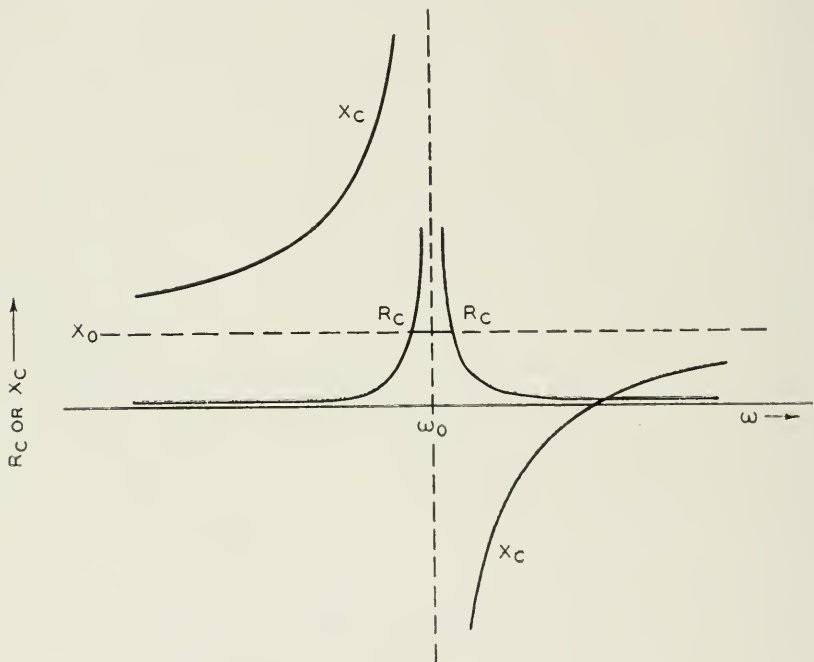
Fig. 34.—A plot of the resistive and reactive components of the impedance $Z_c = R_c + jX_c$, looking into a non-oscillating magnetron through its output circuit [terminals CD of Fig. 31 (a)]. Note the symmetry of the reactance about the point (ω_o, X_o) determining the value of X_o, the loop reactance.

mined by the fact that the reactance curve is antisymmetric about the point (X_0, ω_0). When X_0 has been determined one may determine the admittance, $\dfrac{1}{Z_c - jX_0}$, at the terminals of the ideal transformer of Fig. 31 (b). In terms of the circuit parameters this admittance is

$$\frac{1}{Z_c - jX_0} = \left(\frac{L}{M}\right)^2 Y_{0c}Q_0 + j2Y_{0c}\left(\frac{L}{M}\right)^2 \frac{\omega - \omega_0}{\omega_0}. \qquad (41)$$

The conductance term, $\left(\dfrac{L}{M}\right)^2 Y_{0c}Q_0$, is independent of frequency and the susceptance term, $2Y_{0c}\left(\dfrac{L}{M}\right)^2 \dfrac{\omega - \omega_0}{\omega_0}$ varies linearly with frequency.   If these quantities are plotted from experimental data as functions of ω, the values of Q_0 and $\left(\dfrac{L}{M}\right)^2 Y_{0c}$ may be determined.   To determine the individual values of the factors in $\left(\dfrac{L}{M}\right)^2 Y_{0c}$ for a given magnetron, it is necessary to calculate or somehow to measure a value for $Y_{0c} = \sqrt{\dfrac{C}{L}} = \omega_0 C = \dfrac{1}{\omega_0 L}$.
When used with ω_0 this yields values for L, C, and M. How this is done will not be explicitly discussed here. From the values of Q_0 and Y_{0c}, values of Q_L , Q_{ext} , and η_c may be calculated for any magnetron load G_L'' by the relations (21), (22), (23), and (25). There are other methods of extracting the Q values from the experimental data than that presented above.

When values of the circuit parameters are available one can calculate the values of G_s and B_s from which G_e and B_e are obtained. From the output power the RF voltage is then calculable by equation (38), and from this and the electronic admittance the in-phase and quadrature components of the RF current may be obtained. Using the circuit efficiency, η_c , now determined as a function of load, one can obtain the dependence of the electronic efficiency, η_e , as a function of load conductance from experimental values of the over-all efficiency, η, measured along a constant frequency contour of a Rieke diagram ($\eta = \eta_e\eta_c$). The plot of Fig. 19 was obtained in this way. One is now in possession of values for each of the parameters upon which the electronics of the magnetron depends and may study the relations between them.

9.4   *Pulling Figure:* The Rieke diagram completely specifies the dependence upon load of the magnetron output power and frequency of operation.   Nevertheless, it is convenient to be able to specify by a single parameter the dependence of operating frequency on load changes. The preceding discussion has shown that the changes in load conductance reflected into the resonant circuit of the magnetron, that is, specified at either the primary or secondary terminals of the ideal transformer of Fig. 31 (b), vary output power only. Further, if G_e and B_e are unrelated, load susceptance changes specified at the same points vary frequency only. Since G_e and B_e are in fact related, constant frequency contours on the $G_L'-B_L'$ plane, as has been seen, are inclined to constant B_L' lines at the angle α. Thus changes in B_L' are more effective by the factor $1/\cos\alpha$ in affecting frequency

than equation (37) would indicate. The quantity calculated for $\dfrac{\Delta\omega}{\Delta B_L'}$ from equation (37) multiplied by the factor $1/\cos\alpha$ is a parameter which specifies the dependence of frequency on load. It may be specified at any value of load admittance Y_L', that is, for any position on the $\vec{r}$ plane. Generally, however, one considers the rate of change of frequency with susceptance at matched load, for which Y_L' is equal to the characteristic admittance Y_0' of the output line at the point in question. Using the parentheses ( )$_0$ to indicate that the quantity enclosed is measured at the match point and incorporating the factor $1/\cos\alpha$, one obtains from (37):

$$\left(\frac{\Delta\omega}{\Delta B_L'}\right)_0 = \frac{1}{2}\left(\frac{M}{L}\right)^2 \frac{\omega_0}{Y_{0c}} \frac{1}{\cos\alpha}. \tag{42}$$

The quantity $\left(\dfrac{\Delta\omega}{\Delta B_L'}\right)_0$ as it stands is not a convenient one to measure. For this reason it is customary to specify the total excursion of frequency, $\Delta f = \dfrac{\Delta\omega}{2\pi}$, resulting from a standard variation in $\Delta B_L'$, namely, that obtained by the total possible phase variation of a standing wave of 1.5 voltage ratio in the line at the point in question. This is equivalent to traversing the $\rho = 0.2$ circle on the reflection coefficient plane, shown on Fig. 33. It can be shown that such a variation of load admittance results in a variation of susceptance of ± 0.41 times the characteristic admittance of the line, Y_0', corresponding to a total susceptance variation of $0.82\, Y_0'$. When determined in this way the total frequency excursion is called the pulling figure, PF. Hence by equation (42)

$$PF = \Delta f = \frac{\Delta\omega}{2\pi} = \frac{1}{2}\frac{2\pi f_0}{2\pi}\left(\frac{M}{L}\right)^2 \frac{.82 Y_0'}{Y_{0c}}\frac{1}{\cos\alpha}$$

$$= 0.41\, f_0 \frac{\left(\dfrac{M}{L}\right)^2 Y_0'}{Y_{0c}}\frac{1}{\cos\alpha}. \tag{43}$$

Since $Y_0' = (G_L')_0$, the quantity $\left(\dfrac{M}{L}\right)^2 Y_0'$ in this expression is recognized as $(G_L'')_0$, the load admittance at the primary terminals of the ideal transformer when the line is matched at the point in question, namely, the secondary terminals. Using this fact and equation (23) the pulling figure is seen to be:

$$PF = 0.41\, f_0 \frac{(G_L'')_0}{Y_{0c}}\frac{1}{\cos\alpha} = \frac{0.41\, f_0}{(Q_{\text{ext}})_0}\frac{1}{\cos\alpha}. \tag{44}$$

Although this equation was derived for a specific point in the equivalent circuit, it is of general validity at any point in the output circuit or load line of the magnetron, provided the quantity $(Q_{ext})_0$ is properly interpreted as the external Q measured at match in the line at the same point.

10. SPECIAL TOPICS

10.1 *Frequency Stabilization:* The degree of stability of the operating frequency of the magnetron to load changes is specified by the external Q, as equation (44) indicates. The external Q, by equation (23), may be increased either by decreasing the load conductance, G_L'', or by increasing the circuit characteristic admittance, Y_{0c}. The first alternative may be accomplished by reduction of the coupling between the load and magnetron resonator system. Although this results in greater frequency stability, it entails a reduction in output power. Increase of the characteristic admittance of the magnetron resonator system, on the other hand, increases the energy storage capacity as indicated by equation (20) without appreciably changing the output power. Frequency stability may be increased in this way either by redesign of the magnetron resonator system or by coupling to it a tuned cavity of high unloaded Q. In the latter case, the degree of stabilization, defined as the ratio of energy stored in the combination of magnetron resonator system and stabilizing cavity to the energy stored in the magnetron resonator system alone, is the factor by which the external Q is increased and the pulling figure decreased. In actual practice the stabilizing cavity may be coupled into one of the magnetron resonator cavities or may be built into the output circuit.

10.2 *Frequency Sensitive Loads:* In the preceding sections it has been seen how the load admittance, among other parameters, determines the frequency at which the magnetron oscillates. If this load admittance is itself a function of frequency, it may be possible for the condition of oscillation to be satisfied at more than one frequency. This fact makes for an uncertainty of operation, which is to be avoided. Should the load fluctuate for example, as it does in many applications, the oscillation may jump discontinuously from one frequency to another. If the magnetron is pulsed, it may in certain circumstances oscillate at different frequencies on successive pulses. A tunable magnetron operating into a frequency sensitive load exhibits periodic gaps in its tuning characteristic in which the magnetron cannot be made to operate.

The discussion here will be limited to the specific type of frequency sensitive load consisting of a long line terminated in an admittance, assumed to be frequency insensitive, which differs from the characteristic admittance of the line. The input admittance of such a line is represented by a reflec-

tion coefficient of amplitude, ρ, depending only on the termination [see equation (30)], and of phase, ϕ, depending only on the frequency, f, and the line length, ℓ. Thus:

$$\phi = \frac{2\pi\ell}{\lambda/2} = \frac{4\pi\ell f}{c},$$

from which

$$f = \frac{c\phi}{4\pi\ell}. \tag{45}$$

This equation expresses the linear relation between frequency and phase for the load, specified by the reflection coefficient, into which the magnetron operates. The heavy dashed lines in Fig. 35 represent this relation for two particular line lengths, ℓ_1 and ℓ_2, with ℓ_2 very much longer than ℓ_1. The difference in line length, $\ell_2 - \ell_1$, corresponds to an input phase difference of many times π radians. For the case in Fig. 35, however, this is chosen as an integral multiple of π so that the curve, plotted for the fundamental period 0 to π, will lie in the same range.

The variation of operating frequency of the magnetron with variable phase of the load reflection coefficient is a periodic function whose amplitude increases with increasing amplitude, ρ, of the reflection coefficient. This function may be determined graphically from a Rieke diagram of the magnetron, like that shown in Fig. 33, by traversing the appropriate circle concentric with the center of the diagram and plotting the frequency of operation against phase. In Fig. 35 are plotted such curves for two values of ρ corresponding to different terminations at the end of the long line.

A more detailed analysis of the condition of oscillation shows that it is possible for the magnetron to oscillate stably at those intersections of the magnetron and load frequency characteristics at which the slope of the load line is greater than the slope of the magnetron characteristic. Thus, as indicated in Fig. 35, oscillation may occur at only one frequency for the line of length ℓ_1 if $\rho = 0.2$ but at two frequencies if $\rho = 0.5$. In the latter case the middle intersection, indicated by an open circle, does not correspond to stable oscillation. For a line of length ℓ_2, on the other hand, two oscillation frequencies are possible at both $\rho = 0.2$ and $\rho = 0.5$. If in either case the line lengths ℓ_1 and ℓ_2 are increased by only an approximate quarter wavelength, corresponding to a phase change of $\pi/2$ radians, the light dashed lines labelled ℓ_1' and ℓ_2' in Fig. 35 represent the load characteristics, and oscillation can occur at only one frequency with ρ equal either to 0.2 or 0.5.

If one considers the relationships depicted in Fig. 35 it becomes clear that there are two critical relationships between ρ and ℓ. The first specifies the

values of ρ and ℓ which if exceeded makes oscillation possible at more than one frequency at all phases of the load reflection coefficient. The second specifies the values of ρ and ℓ which must not be exceeded if oscillation is to be possible at only one frequency for all phases of the load reflection coefficient. This latter relation between ρ and ℓ is that for which the slope of the load line is equal to the slope of the magnetron characteristic at its point of inflection.

From what has been said it would appear that the use of long load lines is to be avoided if at all possible. If the magnetron is pulsed and the line
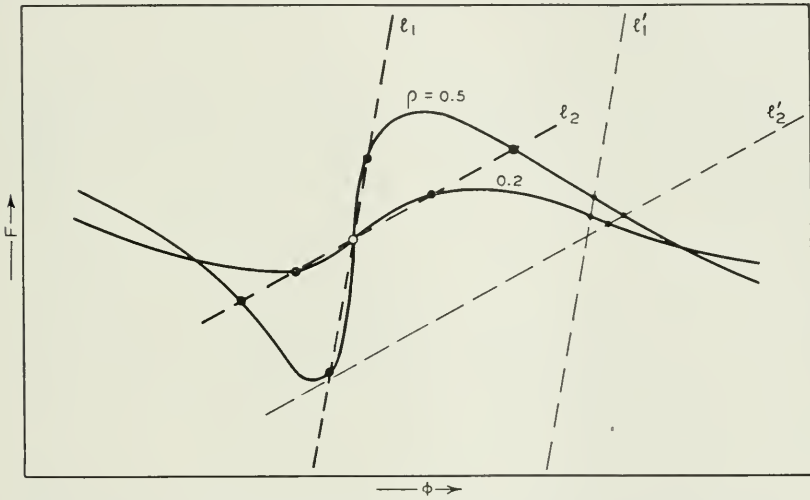
Fig. 35.—Plots of the frequency characteristics of a magnetron and a frequency sensitive load for two magnitudes of reflection coefficient, ρ, and two line lengths, ℓ. The ordinates are frequency and the abscissas are phase angle of the reflection coefficient in the fundamental period 0 to π radians. Stable oscillation occurs at the intersections indicated by filled circles. The open circle indicates a point of unstable operation for the three conditions $\rho = 0.5$ and $\ell = \ell_1$, $\rho = 0.5$ and $\ell = \ell_2$, and $\rho = 0.2$ and $\ell = \ell_2$. In addition, it indicates a point of stable operation for $\rho = 0.2$ and $\ell = \ell_2$. These four points are coalesced on the figure for simplicity. The lines ℓ_1' and ℓ_2' indicate how stable operation at both $\rho = 0.2$ and 0.5 may be attained by an increase of line length of approximately a quarter wavelength.

is sufficiently long, however, it is possible for the oscillation in a pulse to have been completed before any energy reflected from the termination has arrived back at the magnetron. Under these circumstances the magnetron operates into an effectively infinite line which presents its characteristic admittance to the magnetron output. For a pulse of one microsecond duration this would require a line length of about 150 meters. Usually such lengths are not possible or are undesirable by virtue of attenuation, and either the line is made sufficiently short or its length critically adjusted or the standing wave on it reduced so as to cause operation to occur at a single

frequency. Studies have been made of the transient conditions prevailing near the beginning of a pulse when, in establishing the steady state, the successive reflections are returning to the magnetron and, as a consequence, its frequency is changing. Except in a limited region on the Rieke diagram where operation is completely uncertain, it has been shown that the magnetron will settle down to operation at one frequency dictated by the phase of the first reflection, even if oscillation at two frequencies by the previous analysis is possible. Gaps in the tuning curve of a tunable magnetron correspond to the periodic traversal of this uncertain region as the frequency is varied and the load reflection coefficient moves around a constant ρ circle on the Rieke diagram.

10.3 *Magnetron Tuning:* To tune the magnetron oscillator it is necessary to vary a susceptance somewhere in its circuit. It has already been observed how variation of load admittance when reflected into the resonator system as a susceptance change results in frequency pulling. Although for other reasons this variation is usually limited by output circuit design to the order of 0.1% of the operating frequency, it could be increased and used in special instances as a means of tuning the magnetron. Similarly the susceptance of a stabilizing cavity coupled into the resonator system may be varied by tuning the cavity. This method in general enables one to tune over a wider range than does variation of load susceptance since the resonator system is usually more tightly coupled to the stabilizing cavity than it is to the load.

The largest tuning ranges have been attained, however, when it has been arranged to vary one of the frequency determining parameters of the magnetron resonator itself. Schemes have been devised which alter primarily either the inductance or the capacitance of the resonant cavities. Variation of the inductance has been found more convenient at the shorter wavelengths and variation of the capacitance easier at longer wavelengths.

Variation of the inductance may be accomplished by the insertion of a conducting pin into each resonator where the RF magnetic lines of force are concentrated. In a system of hole and slot type resonators it is arranged to move the pins in and out along or near the axes of the holes. Such an arrangement is shown schematically in Fig. 36 (a). As the pins are inserted they reduce the volume available for the magnetic flux, thus reducing the inductance and increasing the frequency. Tuning ranges as great as $\pm 7\%$ of the mean frequency have been attained by this means. In spite of the fact that the ratio of strap to resonator capacitance remains nearly constant, the separation of mode frequencies, as might be expected, decreases with increasing frequency because of the increase in strap inductance resulting from increase of its electrical length. The effect is not so large,

however, that with reasonable tightness of strapping difficult is encountered with mode frequency separation.

Variation of the capacitance may be accomplished by moving a member in the vicinity of the region of large distributed capacitance of the resonators. One such tuning scheme is that shown in Fig. 36 (b) in which a member shaped like a "cookie cutter" is moved in and out of annular grooves in one end of the resonator block, the other end of which is usually strapped. As the member is inserted the mode frequencies decrease. The frequency range available for π mode operation is limited by the fact that the fre-
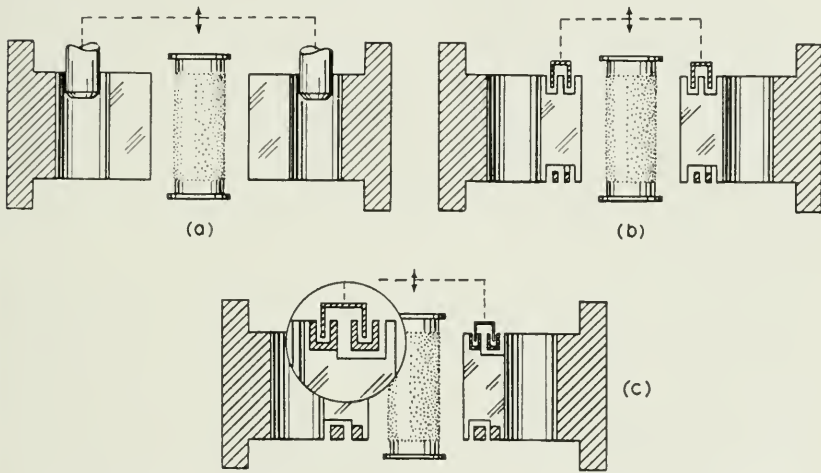
Fig. 36.—Schematic diagrams showing three types of tuning schemes which have been used in magnetron resonator systems. The views represent sections on a diametral plane through the anode structure. The cathode is shown in the center of each part of the figure. The resonators are of the hole and slot type. (*a*) shows the scheme involving variation of the inductance by means of tuning pins, (*b*), the scheme involving variation of resonator capacitance, and (*c*), that involving variation of strap capacitance. Each of the last two schemes is accomplished with a "cookie cutter" shaped member.

quencies of modes of smaller n, which are normally higher than that of the π mode, change faster and eventually cross the π mode frequency.

To understand what determines the frequency of resonance in any mode, it is sufficient to consider the inductances and capacitances present in the average sector of the resonator system across which there is a half wavelength azimuthal variation of RF potential with an extremum at either end. In the π mode of an $N = 8$ resonator system, one may thus consider one eighth of the system, namely, a single resonator; in the $n = 1$ mode, one may consider half of the resonator system. In the case of the half wavelength sector for the $n = 1$ mode, across which the potential varies monotonically (see Fig. 23), the four resonators may be considered as connected essen-

tially in series, making up an equivalent resonator oscillating at the $n = 1$ mode frequency. If one considers further, for the sake of argument, that there is no coupling between resonators, all the modes would have the same frequency but the net inductance and capacitance of the equivalent resonator for each mode of periodicity n, would be proportional to $N/2n$ and $2n/N$, respectively. Thus for the $n = 1$ mode, the equivalent L is four times and the equivalent C is one quarter of the respective values for the π mode. The tuner capacitance added to the equivalent resonator for the $n = 1$ mode, on the other hand, may be considered to be approximately a series-parallel arrangement of capacitances each of which is that between the tuner and one anode segment, C_T: the two parallel capacitances at the positive anode segments connected in series through the tuner ring with the two parallel capacitances at the negative anode segments, the combination having a net capacitance C_T. For the π mode, the net tuner capacitance added per half wavelength potential variation is made up of two capacitances each of magnitude one half of that between the tuner and one anode segment, connected in series through the tuner; the net tuner capacitance is thus $C_T/4$.

By similar reasoning for any mode, one may conclude that the added tuner capacitance per half wavelength sector increases as n decreases, while the net resonator capacitance across which the tuner is shunted decreases as n decreases. In this way the increased effectiveness of the tuner in varying the mode frequencies of the low periodicity modes is accounted for. In actuality the resonators in the half wavelength variation of RF potential are not in simple series connection because of the phase relations between them; the approximation improves for smaller n. Also because of the phase relation between adjacent resonators, the adjacent tuner to segment capacitances are not charged to the same potential and so are not actually in simple parallel connection. Furthermore, the coupling between resonators is important. These considerations modify the above argument somewhat but do not affect the conclusions reached as to the trend of tuning for the different modes. In all the tuning schemes described, second order effects come in through change in the electrical lengths of the straps and tuner as the frequency is varied.

If one arranges to vary the characteristics of the straps by means of a movable tuning member as in the scheme of Fig. 36 (c), considerably greater tuning ranges may be achieved than by the means just described. In this instance, the straps are enlarged to channels of U-shaped cross section, in and out of which the tuning member of "cookie cutter" shape is driven. The effect of insertion of the tuning member is to increase the capacitance per unit length of the strap system by increasing the interstrap capacitance and to decrease the inductance per unit length by effectively increasing the

cross sectional area of the straps. The effects of these changes upon the mode frequencies may be seen from the considerations of the effect of straps on the mode frequencies of the unstrapped resonator system already discussed. The increase in strap capacitance increases the wavelength of the π mode; the decrease in strap inductance decreases the wavelength of modes of smaller n. As the tuning member is inserted, the mode frequencies separate, and no limitation on the range of operation in the π mode is imposed by interference from other modes. This means has been used for tuning ranges of better than $\pm 6\%$, but there is nothing inherent in the scheme to prevent its use for ranges considerably in excess of this value.

Tuning of the magnetron resonator system by any of the means described above alters its characteristic admittance, $Y_{0c} = \sqrt{\dfrac{C}{L}}$, and hence the stored energy. For fixed output coupling and load admittance this amounts to a variation of the effective loading as specified by the external Q. In some cases, attempts have been made to compensate for this by designing into the output circuit a frequency characteristic which keeps the external Q, and hence the pulling figure, more nearly independent of frequency.

Each of the tuning schemes described above is adaptable to precise and specific frequency adjustment or to frequency variation at a slow rate. For tuning over small ranges it is possible to vary the magnetron frequency electronically, enabling one to frequency modulate its output at a high rate. Of the schemes tried for this purpose there may be mentioned that in which an intensity modulated electron beam of a specific velocity is shot parallel to a superposed DC magnetic field through one of the magnetron resonators or a closely coupled auxiliary cavity.

Concerning the variation of magnetron operating frequency will be mentioned finally the shifts which are brought about by temperature variations of the resonator block. Since the resonator system is generally constructed entirely of copper, it expands or contracts uniformly with temperature and the frequency shifts are those attained by a uniform scaling of the resonator system by a very small factor. The temperature coefficient of frequency may readily be seen to equal the negative of the linear coefficient of thermal expansion of copper.

10.4 *Electronic Effects on Frequency:* In Section 3.6 *Induction by the Space Charge Cloud* it has been seen that the electrons moving in the interaction space of the magnetron oscillator contribute an admittance Y_e connected to the resonator system. The magnitude of the negative electronic conductance determines the energy delivered to the circuit and thus the amplitude of oscillation. The electronic susceptance, arising from the phase relation between the space charge spokes and the maximum of the tangential

retarding field, affects the frequency of oscillation. Under equilibrium conditions the magnitude of the current, I_{RF}, induced in the anode segments, its phase relative to the RF voltage between the segments, and the operating frequency adjust themselves such that this admittance, $Y_e = I_{RF}/V_{RF}$, equals the circuit admittance Y_s. The induced RF current and its phase relative to the RF voltage both depend upon the parameters such as V and B, governing the electronic operation of the magnetron. The frequency change at constant load arising from changes in V or B when divided by the change in the DC current, I, drawn by the magnetron, is called the frequency "pushing" and is measured in mc/s per ampere.

A further effect of the electronic susceptance is the shift of the resonant frequency between the oscillating and non-oscillating conditions of the magnetron. In general the oscillating frequency is lower than the resonant frequency of the non-oscillating magnetron. Thus the electronic susceptance is capacitive with the space charge spokes moving somewhat ahead of the field maxima during oscillation. This shift in the resonant frequency is important in pulsed radar systems where the same antenna is used for both receiving and transmitting. An echo of the transmitted pulse on its return then encounters a high, off-resonance impedance at the magnetron which absorbs very little of the returned energy. Most of the received pulse energy is consequently made available to the receiver. For some magnetrons the shift off resonance is not sufficient and other means such as the use of the so-called ATR box are required to divert the received pulse energy into the receiver.

10.5 *Frequency Spectrum of a Pulsed Magnetron:* Only if a generator operates for an infinitely long time is its output "monochromatic", that is of a single frequency. The period of operation of a CW oscillator is generally long enough to make any deviations from this unobservable.

If, however, the oscillation is modulated as in the case of pulsed magnetrons for which the pulse duration is of the order of one microsecond, it is readily detectable that the output is "polychromatic" with the energy distributed throughout a band of frequencies. The plot of the distribution in frequency of the energy generated is called the frequency spectrum.

This state of affairs is perhaps made plausible if one considers pulsing the magnetron to be a very drastic means of amplitude modulating its output. Already, in connection with the case of two coupled circuits, it has been seen how amplitude modulation of an oscillating system in time has associated with it the distribution of the energy over more than one frequency. In an analogous but more complicated manner, a sinusoidal oscillation which is amplitude modulated by a nonsinusoidal pulse shape like that of Fig. 37 (a) is compounded of frequencies not now discrete but

distributed continuously throughout a band, Fig. 37 (a'). The energy distributions in time and frequency are related mathematically by the Fourier transform.

The breadth of the frequency distribution is inversely proportional to the modulating pulse width as shown in Fig. 37. The spectrum widths of operating magnetrons may exceed the theoretical width by a factor of not more than two for reasons not altogether clear.

10.6 *Oscillation Buildup—Starting:* Of importance in the design and operation of pulsed magnetrons are the phenomena associated with the buildup of oscillation when the voltage is applied. Here will be discussed briefly what is known about the problem, the factors upon which the rate of buildup of oscillation in the circuit depends, how this is related to the rate
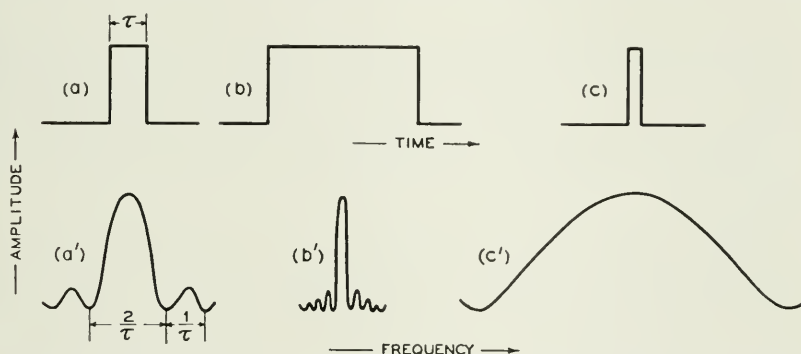
Fig. 37.—Plots of DC voltage pulse shapes (a), (b), and (c) and the frequency spectra (a'), (b'), and (c') resulting from each, respectively. These are shown for the idealized case of perfectly rectangular voltage pulse shapes.

of voltage buildup and to pulser characteristics, and how these factors result in various types of magnetron starting behavior.

The rate of buildup of oscillation in the magnetron depends not only upon the circuit characteristics but also upon the electronic behavior with increasing DC and RF voltages. Because of the nonlinear nature of the electron interaction, the amount of energy fed to the circuit per cycle is dependent upon the amplitude of the oscillation. Considering the condition for oscillation given by the first of equations (35) and expanding it to include during buildup a term to account for the energy being stored in the resonators, one may derive a simple expression for the rate of increase of RF voltage amplitude. The energy stored in the circuit at any instant being

$W = CV_{RF}^2$, the rate at which energy is being stored is $\dfrac{dW}{dt} = 2CV_{RF}\dfrac{dV_{RF}}{dt}$.

This corresponds to a conductance term, G, obtained by equating $\dfrac{dW}{dt}$ and GV_{RF}^2 :

$$G = 2C \frac{dV_{RF}}{dt} \frac{1}{V_{RF}}. \tag{46}$$

The necessary condition for oscillation during buildup is thus:

$$G_e(V_{RF}) + G_s + G = 0, \tag{47}$$

from which one obtains

$$G_e(V_{RF}) + G_s + 2\frac{C}{V_{RF}}\frac{dV_{RF}}{dt} = 0. \tag{48}$$

Since oscillation in the magnetron does in fact build up, $|G_e(V_{RF})|$ must be greater than G_s when oscillation starts. When equilibrium is reached, $G_e(V_{RF}) + G_s = 0$. Thus $|G_e(V_{RF})|$ must be a decreasing function of V_{RF} crossing the value G_s at the operating point. By equation (48), V_{RF} thus builds up rapidly at first and then more slowly as $|G_e(V_{RF})|$ approaches G_s.

Increase in load, resulting in an increase of G_s, decreases the rate of oscillation buildup.[22] The dependence on the total resonator capacitance indicates that for magnetrons of different sizes related by a simple scale factor, G_e and G_s under such scaling presumably remaining invariant, the rate of buildup decreases with increasing wavelength. Unknown in relation (48) is, to be sure, the exact transient dependence of G_e on RF and DC voltages, magnetic field, and interaction space geometry. It is known, however, that an increase of cathode diameter, although it is accompanied by a decrease in electronic efficiency, does reduce the difficulty with "moding" resulting from failure to start in the π mode.

Associated with the rate of buildup of RF oscillation in determining magnetron starting behavior is the rate at which the DC voltage is applied. From the discussion of the electronics of the magnetron it is clear that oscillation is not possible at all values of V but only for a limited range near that which provides synchronism between the electron motion and the rotating field pattern. If the pulser can apply a voltage which rises to a value in this region and which can remain at substantially this value regardless of the current drawn, no difficulty is encountered. However, should the pulser regulation be such that to secure operation at a given voltage

---

[22] This is directly the opposite of the behavior with respect to load variations of a circuit being driven in such a way that a constant amount of energy is fed in per cycle, in which case the rate of buildup is inversely proportional to Q.

and current it is necessary to apply a voltage which on no load would rise to a value considerably higher than the operating value, the rate at which the voltage passes through the range of possible operating values and the relation of this rate to that of RF buildup are extremely important. Should the pulse voltage rise so rapidly as to pass through the region where oscillation is possible before the RF oscillation can build up and cause the magnetron to pass current, which by modulator regulation keeps the DC voltage from rising further, the magnetron fails to start. Clearly, the more rapid the oscillation buildup the more rapid a voltage rise is permissible. Conversely, for a given rate of DC voltage rise, failure to start should appear at greater load and longer wavelength as relation (48) implies. Experience has corroborated both of these conclusions. It is also clear by equation (48) that the equalization of loading of the doublet modes of the same periodicity, which is achieved by proper location of strap asymmetries, equalizes their starting times and makes possible interference with π mode starting less likely.

When oscillation in the π mode fails, the magnetron may fail to oscillate at all or may oscillate in another mode for which the operating voltage in a harmonic is higher than but close to that of the π mode. In this case, as has been seen, $\dfrac{2\pi f'}{|k'|} > \dfrac{2\pi f}{N/2}$, and the Hartree line of the "second" or "primed" mode lies just above that of the π mode. This case in which oscillation in the π mode is skipped for oscillation in another mode represents the most common type of "moding" encountered in pulsed magnetrons. If, on the other hand, $\dfrac{2\pi f'}{|k'|} < \dfrac{2\pi f}{N/2}$, and the Hartree line of the harmonic of the "second" mode lies just below that of the π mode, as in Fig. 16, the magnetron is observed to oscillate first in the "second" mode before oscillation at low currents in the π mode commences. When the mode driven during the interval at the top of the pulse is the π mode, oscillation in the "second" mode occurs only momentarily on the rise and fall of each pulse as the voltage passes through the range of operating values for this mode.

The starting behavior of pulsed magnetrons may be shown by the so-called dynamic performance chart or *V-I* plot on which the course in time of the voltage and current is shown. In Fig. 38 are shown three *V-I* plots of this type. The initial current rise is the charging current of the cathode to anode capacitance. The current rise when oscillation commences is very rapid and is shown as a dashed line. After remaining for the major part of the pulse at the operating point, indicated in Fig. 38 by a large dot, the current and voltage fall during which they follow closely a constant *B* line of the static performance chart (see Fig. 17).

In Fig. 38 (a) is shown the dynamic *V-I* plot for normal operation in the

π mode. In Fig. 38 (b) is shown a dynamic plot after the voltage control has been raised above the point of π mode failure. Here the magnetron does not oscillate at all. As seen in Fig. 38 (c), further increase of the voltage control of the pulser makes possible oscillation in the $k = -5$ ($p = -1$) harmonic of the $n = 3$ mode ($N = 8$). In both (b) and (c) of Fig. 38 it is of interest to note how the magnetron tries to oscillate in the π mode as the voltage at the end of the pulse falls through the range of permissible values. These attempts at oscillation are indicated by the magnetron drawing in this region small amounts of current which vary from pulse to pulse.
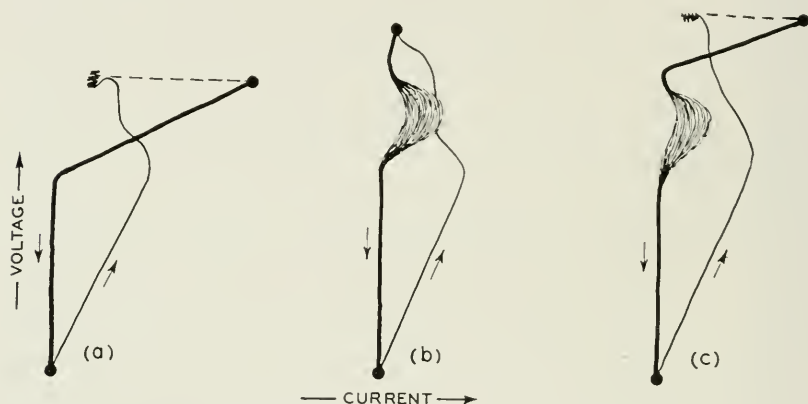
Fig. 38.—Three so-called dynamic V-I plots or dynamic performance charts illustrating the mode skip phenomenon. The plots are copies of presentations obtained with an oscilloscope whose vertical deflection is proportional to DC voltage and whose horizontal deflection is proportional to DC current. On any plot the heaviness of the lines is roughly inversely proportional to the rate at which the coordinates are traversed, and the large dot at the upper extremities represents the operating point at the top of the pulse. (a) shows shows normal operation in the π mode. (b) shows failure to oscillate in any mode, the π mode being skipped. (c) shows oscillation in a harmonic of a mode of smaller n.

10.7 *Magnetron Cathodes:* One important component part of the magnetron oscillator which to this point has not been discussed in detail, but which has been assumed present and operating satisfactorily, is the cathode. Its duty is to supply the electrons which serve as the intermediaries between the DC and RF fields. In terms of the usual requirements of vacuum tube cathodes, the number of electrons demanded of a magnetron cathode is little short of prodigious. Magnetron cathodes may be required to deliver current densities of the order of 50 amperes per square centimeter as contrasted with the 0.5 amperes per square centimeter emitted by oxide cathodes in high vacuum tubes normally.

How oxide surfaces are capable of emitting such enormous currents is

not yet entirely clear.   It is certain, however, that the RF oscillation in the magnetron is responsible for the fact that such currents may be attained. This is made evident experimentally by measuring the current that may be drawn when the magnetic field is reduced so that oscillation is impossible and comparing it with the current drawn during oscillation.   At temperatures even in excess of the operating temperature such pulsed emission currents of excellent magnetrons may run as low as 1% of the currents flowing during oscillation.   Furthermore, the fields at the cathode during these measurements may actually exceed those present during oscillation because of the absence of the dense space charge clouds in the interaction space.

In the process of phase selection of electrons, as previously discussed, those electrons starting from the cathode in a phase such as to gain energy from the RF field are removed from the interaction space and driven into the cathode.   They impart to the cathode the energy they have gained, causing secondary electrons to be emitted and the cathode temperature to increase.   In most magnetrons the amount of this returned energy is about 5% of the input, though in some cases it may become as high as 10%.   This means that the average energy of back bombardment may run as high as 75 watts per square centimeter of cathode surface.   This amount of energy must be dissipated under equilibrium conditions by radiation and conduction.   In general, the short wavelength magnetrons are run with no heater power supplied to the cathode, the cathode temperature being maintained by back bombardment.   In many cases, cathode overheating by this process is actually the limitation on the operating capabilities of the magnetron.

There is some experimental evidence that the large current drawn from the magnetron cathode is not primarily made up of secondary electrons but may result from an "enhanced" primary emission.   This is supported by the fact that the secondary electrons emitted at the return to the cathode of the "out of phase" electrons are themselves, assuming negligible emission time, largely "out of phase" electrons later to return to the cathode.   However, the emission of large numbers of secondary electrons may lead to an "enhanced" primary emission by a process not now understood.   As possible processes may be mentioned field emission or an actual lowering of the cathode work function, each brought about by the fields in the cathode coating which result from the charge loss attendant upon the secondary emission.   Ionic conduction or electrolytic action in the coating may also contribute in some manner to a lowered work function and to the "enhanced" emission, although such ionic processes would generally involve time intervals longer than a microsecond.   The actual mechanism involved, however, is still speculative.

A word about the relation of magnetron scaling to magnetron cathode

problems will here be in order. It has been seen earlier that, in scaling all magnetron dimensions by a factor α, the magnetic field changes by a factor $\frac{1}{\alpha}$ while the current and voltage remain unchanged. That this places severe requirements on the cathode may be seen by considering the scaling of a 10 cm. magnetron down to 1 cm. The operating current may be 20 amperes in both cases. If this corresponds to a current density of 5 amperes per sq. cm. at 10 cm., 500 amperes per sq.cm. will be required at 1 cm. What is more, the back bombardment in watts per sq. cm. is increased by a factor of 100. Both of these requirements are completely unreasonable and preclude direct scaling in this instance. Consequently, an attempt is made in such cases to decrease the current density by increasing both the cathode length and diameter. Increasing the latter usually involves increasing the anode diameter and the number of resonators. Even so, current densities may exceed 50 amperes per sq. cm., as stated earlier.

The pulsed magnetron cathode at the shorter wavelengths (less than 10 cm.) in the centimeter band is a limiting factor in magnetron design. In CW magnetrons the small size of the interaction space has made the cathode an important and difficult design problem throughout the centimeter wavelength region. Considerable effort has been expended in a number of laboratories not only to understand the physics of the operation of the magnetron cathode but to find suitable materials and constructional designs. It has become clear that a good magnetron cathode which will meet the special conditions of high current density and high voltage gradient and the considerable electron back bombardment are these: (1) sufficient primary emission to enable the magnetron to start and to supply a part of the required operating current; (2) sufficient secondary emission to supply the remainder (it may be practically all) of the required current density through whatever mechanism is involved; (3) sufficient active material to permit satisfactory life; (4) some mechanical means of holding the active material on the cathode surface; (5) sufficiently low electrical resistance of the coating to permit large bursts of current without undue local heating and high back bombardment without excessive coating temperature; and (6) satisfactory over-all heat dissipation characteristics, conductive and/or radiative, to keep evaporation of active material to a minimum.

In pulsed magnetrons of wavelength 10 cm. or greater it has generally been possible to use plain oxide coated cathodes. Nickel base material is generally used and the active material is the usual double carbonate coating (reduced to the oxides during activation). At wavelengths of 3 cm. and shorter the development of satisfactory pulsed magnetrons would have been impossible without the development of special cathodes. In the main,

these have been aimed at meeting requirements (3) to (5) above. The constructions have made use of wire meshes and of sintered nickel matrices both to reduce the coating resistance and to hold sufficient material on the cathode in a manner such that it may be dispensed gradually during life.

10.8 *The Magnetic Circuit:* The magnetic field required for operation of the magnetron oscillator is generally obtained, except in laboratory experiments, by means of a permanent magnet. At long wavelengths and in early models at shorter wavelengths, the magnetron and permanent magnet are separable. Building the magnetic pole faces into the magnetron structure itself and attaching the magnetic material to it has made possible the reduction of the over-all magnet gap, and hence total magnet weight, as well as the use of mechanically superior axial cathode mountings. The resulting so-called "packaged" magnetron design has been used at shorter wavelengths where the magnetic fields are high but need not extend over a large area. The total magnet weight under these conditions is much less than that required in a separate magnet. Needless to say, the possession of good permanent magnet material and the work done on magnet design have contributed materially to the success of the centimeter wave magnetron.

10.9 *Magnetron Measurements:* The fundamental measurements made on the magnetron oscillator have already been discussed or alluded to where the performance characteristics of the magnetron and its circuit theory are described. Here will be described briefly the technique of measurement. Magnetron measurements are of two general types. One is made on the oscillating magnetron and the other on the non-oscillating magnetron. The latter may be made at any stage in the fabrication of the magnetron after its anode structure and output circuit are completed. Figs. 39 and 40 illustrate schematically the apparatus employed in these tests.

Perhaps the best way of describing the techniques of magnetron measurements is to list all of the parameters, quantities, or characteristics associated with such measurements and for each to give the definition, method of measurement or calculation, or the way it is put together from other data, as the case may be. In any event, the list given below permits of ready reference. Although the text applies directly to pulsed magnetrons the simplifications for CW magnetrons are obvious.

The *DC magnetic field*, B, in which the magnetron operates is generally supplied in the laboratory by an electromagnet, the field in the gap generally being calibrated in terms of the current passed through the magnet coils.

The *peak DC voltage*, V, applied to the magnetron cathode is measured by means of a peak voltmeter or by observing a known fraction of the voltage

pulse on the calibrated screen of an oscilloscope. In a simple but suitable peak voltmeter it is arranged to charge a condenser through a diode to the peak voltage which may then be measured with a high resistance DC voltmeter.

The *peak DC current*, *I*, drawn by the magnetron is measured by passing the current through a known resistance, usually one or two ohms, and determining the peak voltage developed across the resistance by means of a peak voltmeter or calibrated oscilloscope.

The *average DC current* drawn by the magnetron is the current measured on a DC meter connected in one leg of the pulsing circuit, as shown in Fig. 39.

The *pulse duration*, *τ*, as its name implies, is the length of the time during which the voltage, usually measured near the top of the pulse, is maintained across the magnetron. It may be determined from the pulse presentation on an oscilloscope having a calibrated sweep or it may be calculated as indicated below when other parameters are known.

The *pulse recurrence rate*, *pps*, is the repetition frequency at which the voltage pulse is applied and is determined by the frequency of the calibrated primary oscillator driving the pulser or modulator circuit.

The *duty cycle*, defined as the fraction of time the pulsed magnetron operates, may be determined as the ratio of average to peak DC current or as the product of the pulse duration and the pulse recurrence rate.

The *peak input power* is the product of the peak DC voltage and the peak DC current.

The *average input power* is the product of the peak input power and the duty cycle.

The *voltage and current pulse shapes* as observed with oscilloscopes are of importance in studying the spectrum and moding characteristics of the magnetron under test.

The *dynamic V-I plot*, or *dynamic performance chart*, is viewed on an oscilloscope in which at any instant the vertical deflection is proportional to peak DC voltage and the horizontal deflection proportional to peak DC current. Three such plots are shown on Fig. 38 and their usefulness is indicated in the corresponding text.

The *average output power* is the average centimeter wave power delivered to the useful load. The simplest and most foolproof method of measuring this power is to absorb the energy in a column of water. From a determination of the rate of water flow and its temperature rise the power may be calculated readily. The water column terminating the coaxial line or wave guide of the test apparatus is made reflectionless either by tapering it or preceding it by a quarter wavelength matching plate of proper dielectric constant, analogous to the optical quarter wave plate.

The *peak output power* may be calculated as the average output power divided by the duty cycle.
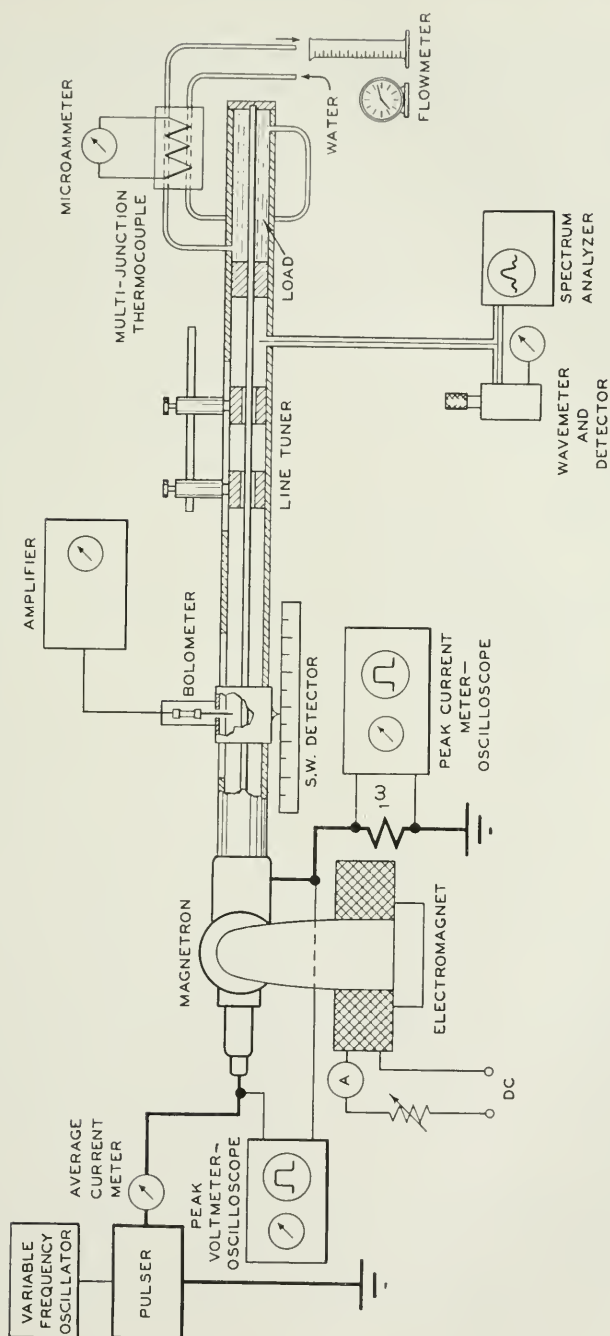
Fig. 39.—A schematic diagram showing the arrangement of apparatus generally used in making measurements on an operating magnetron in the laboratory.

The *over-all efficiency* of operation is the ratio of the peak output to peak input powers or the ratio of average output to average input powers.

The *frequency of oscillation* of the magnetron is determined by feeding a small amount of the RF power into a calibrated variable frequency resonant cavity of high Q and by means of a detector observing the frequency at which the cavity absorbs or passes power.

The *load impedance* into which the magnetron operates is determined by measurement in the output line of the voltage standing wave and its phase with respect to some previously chosen reference point. This measurement is made with a *standing wave detector* in which it is arranged to move an electrostatic probe along a section of slotted line in which it samples the RF energy in the line. The energy picked up by the probe is detected after sufficient attenuation either by a crystal or by a small bolometer whose resistance is a function of the energy fed into it. The voltage standing wave
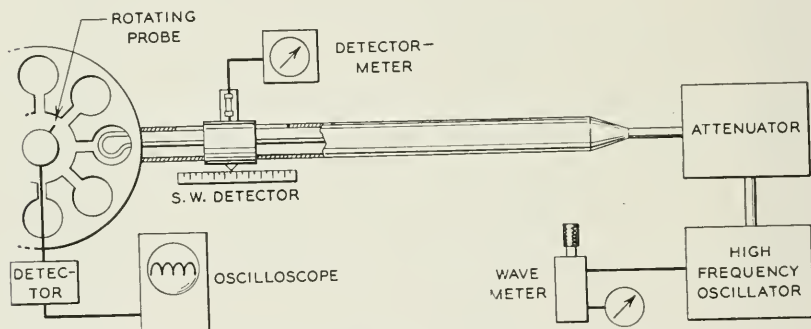
Fig. 40.—A schematic diagram showing the arrangement of apparatus used in making measurements on a non-oscillating magnetron or magnetron resonator block.

ratio and its phase may be translated into impedance, admittance, or reflection coefficient in the manner already discussed. The magnetron load impedance may be varied from the match presented by the terminating water column by means of a *line tuning section* or *tuner*. In one such line tuning section, two sleeves constituting in effect quarter wavelength lines of low characteristic impedance may be moved relative to one another to vary the magnitude, and moved together to vary the phase of the standing wave.

The *Rieke diagram*, is a plot of constant output power and frequency on a reflection coefficient plane. Its construction thus involves the measurement of output power, frequency, and load impedance as the line tuners are moved over a wide range of positions. The operating parameter maintained constant is usually chosen to be the peak DC current. Such a diagram is shown in Fig. 33.

The *pulling figure*, *PF*, defined as the maximum frequency excursion of the magnetron as the load reflection coefficient traverses the $\rho = 0.2$ circle, may be obtained from the Rieke diagram. It may be measured directly by two simple wavemeter measurements taken at the frequency extrema occuring as the standing wave of 1.5 voltage ratio is moved up and down the line. Various types of *standing wave introducers* have been devised to produce a reflection coefficient of $\rho = 0.2$ for the specific measurement of pulling figure.

The *performance chart* is a plot of constant magnetic field, peak output power and over-all efficiency contours on a *V-I* plane. Its construction thus involves the measurement of peak DC voltage, peak DC current, and peak output power at several magnetic fields. Sometimes frequency, pushing figure, and spectrum appearance are also determined. Figs. 17 and 20 are examples of this chart.

The *frequency spectrum* is the distribution of energy with frequency for a pulsed magnetron and is displayed on a so-called *spectrum analyzer*. This analyzer is a very narrow band tunable radio receiver whose pass band is varied periodically twenty or so times per second over several *mc/s*. The response of the receiver to the frequency distribution of energy appears on an oscilloscope whose sweep is synchronized with the pass band frequency variation.

The *pushing figure* is the instantaneous frequency change in *mc/s* per ampere change in peak DC current at constant load. It may be obtained by measuring the frequency shift on a spectrum analyzer as the pulse current is changed by a known amount. The current change must be executed rapidly enough to avoid frequency shifts arising from temperature changes.

Impedance measurements on the non-oscillating magnetron involve the use of a variable frequency RF oscillator feeding power through an attenuator and a standing wave detector into the output circuit of the magnetron (see Fig. 40). These measurements determine as a function of frequency the impedance Z_c discussed in the text.

Mode frequencies are determined from the impedance measurements on the non-oscillating magnetron by noting the frequencies at which the input standing wave is observed to go through a minimum. They may also be determined by the observation of energy maxima with a pickup loop or probe placed in the resonator system.

Mode identification is made by observing the periodicity of RF field in the interaction space of the resonator system. This is done by sampling the field with a rotating RF probe placed in an axial cylinder corresponding to the cathode as shown in Fig. 40. The probe response is detected and displayed on an oscilloscope with sweep synchronized to the rotation of the probe.

PART II

DEVELOPMENTAL WORK ON THE MAGNETRON OSCILLATOR AT THE BELL TELEPHONE LABORATORIES, 1940–1945

11. General Remarks

IN THE first part of this paper the fundamentals of the theory of the magnetron oscillator have been discussed. The objective has been to establish for the reader a general picture of the nature of the electronic mechanism and of the role played by the RF circuit and load.

In the second part of the paper is traced the research and development work done at the Bell Telephone Laboratories on the magnetron oscillator during the war years, 1940–1945. The effort was directed, for the most part, toward the development of magnetrons to meet definite radar needs.

Fifteen different types or families of magnetrons were developed at the Bell Laboratories during the war. Included among these are some 75 separate Western Electric Company or RMA code numbers. It has been found most convenient to discuss the work done on each type of magnetron as a unit, although, to be sure, there has been considerable interplay between projects. Something is said of the origin of each type, of the problems encountered in its development, and of the solution of these problems, in some cases involving studies and experiments of general interest. Special characteristics and general performance data for each type of magnetron discussed are given. Included also is a general discussion of the work done on magnetron cathodes, which, although carried out on specific magnetrons, has been of general applicability to all.

Before proceeding with the detailed discussion, it would be well for the reader to recognize the general scope of the work to be described and the general nature of the problems encountered. The work of the Bell Laboratories in the development of pulsed magnetrons for radar use has extended over practically the whole range of effort surveyed in the INTRODUCTION. Work has been done throughout the range of wavelengths from 45 cm. to 1 cm. and on magnetrons capable of developing over one megawatt peak RF power. It has included work on such features as tuning, coaxial and wave guide outputs, several types of resonator systems and strapping schemes, and on the incorporation of the magnetic circuit into the magnetron structure in so-called "packaged" types.

The scope of the developments to be described may be judged from Figs. 41, 42, 43, and 44. That part of a magnetron oscillator which perhaps best gives one an idea of size and wavelength range is the resonator block. In Fig. 41 is shown a series of resonator blocks ranging in resonant fre-
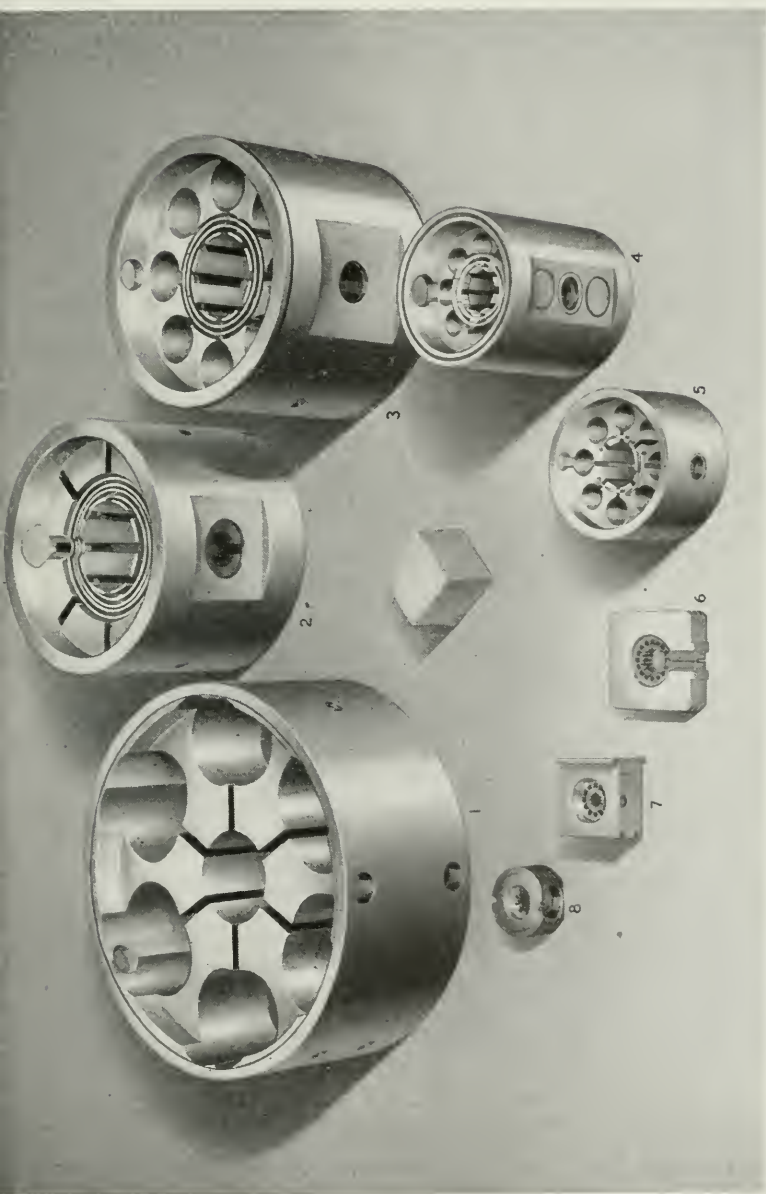
Fig. 41—Resonator blocks of centimeter wave magnetrons developed at the Bell Telephone Laboratories. From the largest to the smallest these resonator systems are those of: (1) the 700A-D magnetrons of fixed frequencies near 700 mc/s; (2) the 5J26 magnetron, tunable over the frequency range 1220 to 1350 mc/s; (3) the 4J21-30 magnetrons of fixed frequencies near 1280 mc/s; (4) the 720A-E magnetrons of fixed frequencies near 2800 mc/s; (5) the 706AV-GY magnetrons of fixed frequencies near 3000 mc/s; (6) the 4J50 magnetron at 9375 mc/s; (7) the 725A magnetron at 9375 mc/s; and (8) the 3J21 magnetron at 24,000 mc/s. Note the hole and slot type resonators of (1), (3), (4), (5), (6), and (7); the slot type resonators of (2); and the vane type resonators of (8). Note the double ring straps in (3), (4), (6), and (7); the double ring channel straps for tuning in (2); the early British type wire strapping in (5); and the unstrapped "rising sun" type structure of (8). A one inch cube has been included in this and other photographs for size reference.
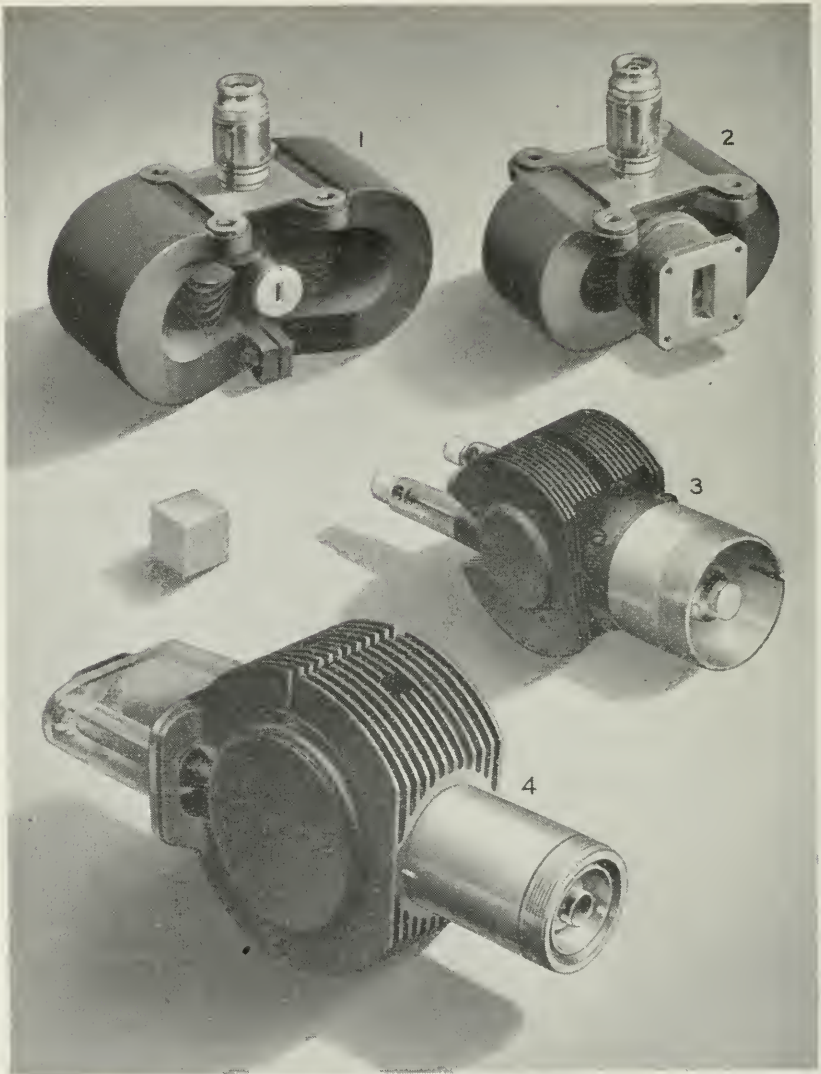
Fig. 42—Four fixed frequency magnetron oscillators.   They are: (1) the 3J21 magnetron (60 kw., 24,000 mc/s); (2) the 4J52 magnetron (100 kw., 9375 mc/s); (3) the 720A-E magnetron (1000 kw., ∼2800 mc/s); (4) the 4J21-30 magnetron (600 kw., ∼1280 mc/s).   Note the two types of output circuit, coaxial and waveguide; the use of packaged magnets; and the two types of input leads.

quency from 700 mc/s to 24,000 mc/s.[23]   In Fig. 42 is shown the external view of four magnetrons of operating frequencies distributed over the range

[23] In this and other photographs of PART II either a one inch cube or a line of length one inch has been included for size reference.
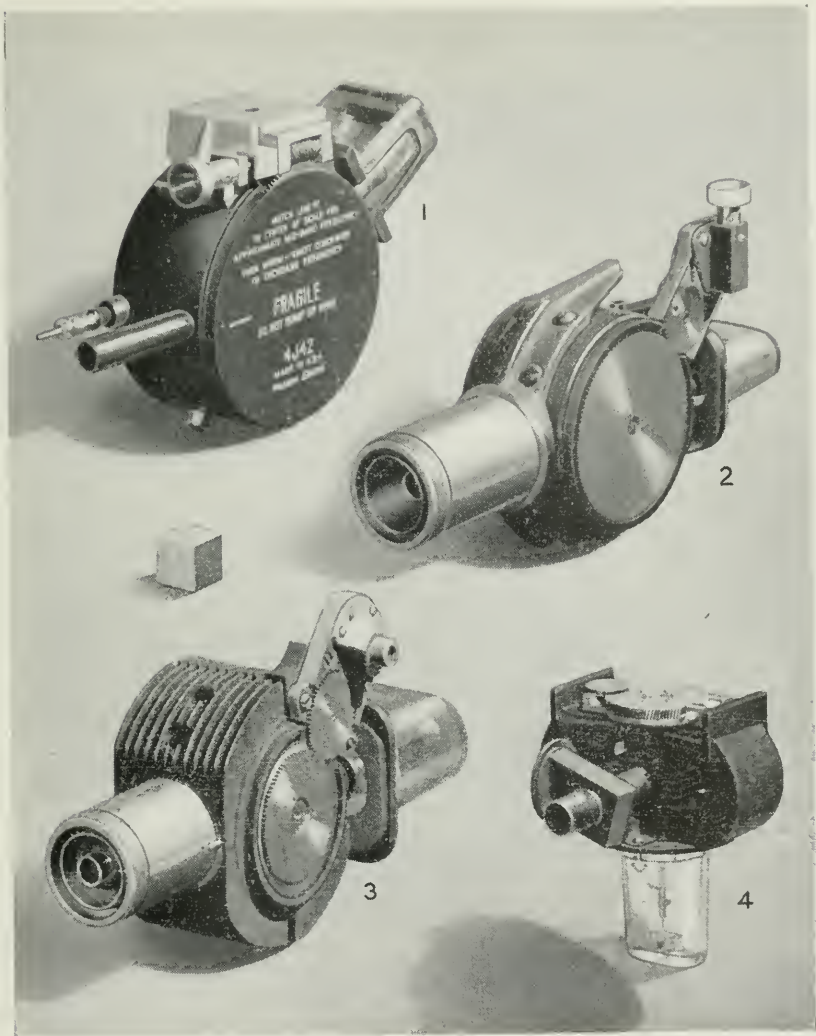
Fig. 43—Four tunable magnetron oscillators. They are: (1) the 4J42 magnetron (40 kw., 660 to 730 mc/s); (2) the 4J51 magnetron (275 kw., 900 to 970 mc/s); (3) the 5J26 magnetron (600 kw., 1220 to 1350 mc/s); and the 2J51 magnetron (55 kw., 8500 to 9600 mc/s).

in which work has been done. It illustrates both the coaxial and wave-guide types of output circuits, the use or not of attached magnets, and the two types of input leads and cathode supports. In Fig. 43 is shown the group of tunable magnetrons developed as replacements for fixed frequency models with which they are electrically and mechanically inter-
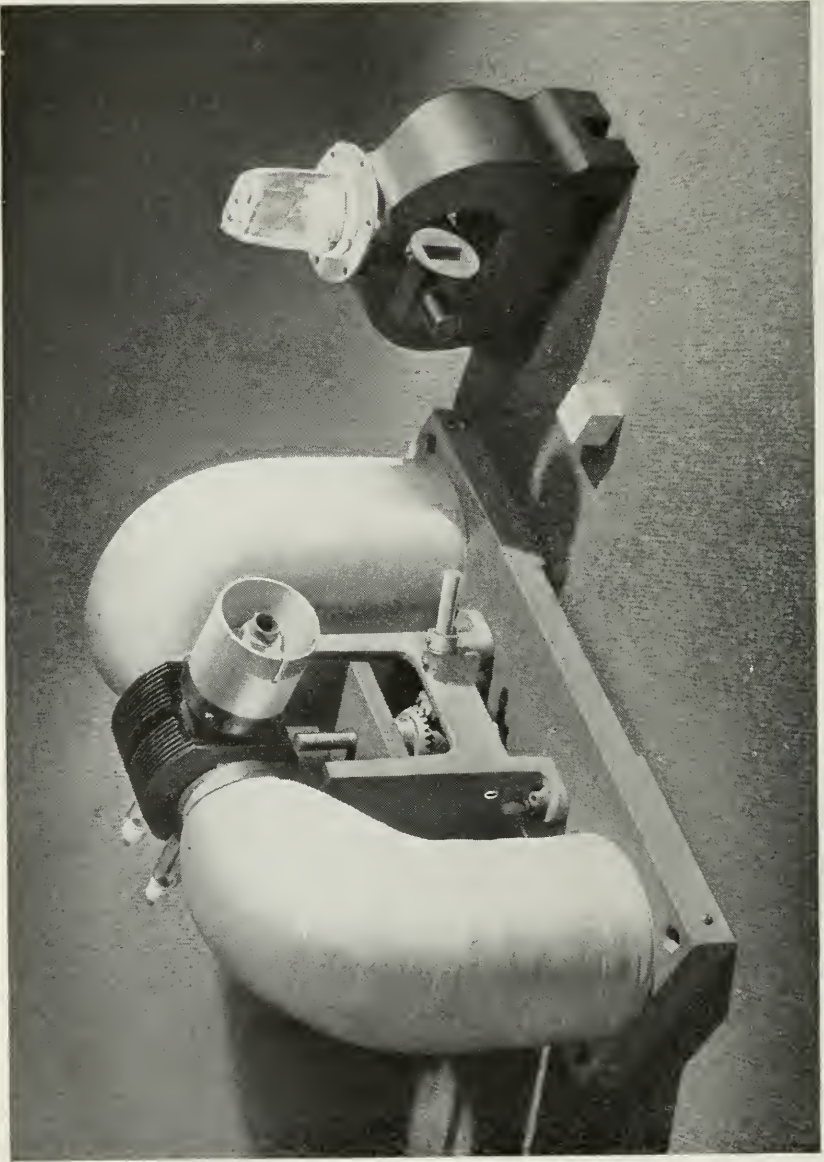
Fig. 44—The 720A-E (1000 kw., ~2800 mc/s) and the 725A (55 kw., 9375 mc/s) magnetrons, shown at the left and right respectively, each mounted in its magnet.

changeable. Finally, in Fig. 44 are shown two magnetron oscillators, the 720A and 725A, mounted in their magnets. By comparison with Fig. 42, the space and weight saved by packaging may be seen. A fair comparison is that between the 725A and magnet of Fig. 44 and the 4J52, number 2 in Fig. 42. Both are 3.2 cm. models, the latter, moreover, being capable of generating higher power.

The designer of a magnetron oscillator is faced with a variety of tasks. If the magnetron is to be used in a specific application he has at his disposal data concerning the amount of power available to drive the magnetron, the nature of the pulsing if such is to be used, the frequency of operation, mechanical features having to do with form and weight, and an idea of what the user hopes or expects to obtain in the way of output power, frequency stability, and operating efficiency. It is the problem of magnetron design to arrange the resonator system, output circuit, cathode, magnetic circuit, and mechanical features to meet these requirements if possible.

In the design of the resonator system it must be arranged to achieve the proper frequency of operation, proper characteristics regarding modes, the proper size of interaction space, and other characteristics which have a bearing on the electronic operation. In special cases a tunable resonator system must be provided.

In the design of the output circuit it is necessary to arrange the type of coupling to the resonator system, the necessary impedance transformation from resonator to load, the type of external coupling, a vacuum seal, and generally to take into account the possibility of electrical breakdown when the power delivered is high. ,

In the design of the magnetron cathode, attention must be paid to its surface and how it is equipped to meet the rigorous demands made of it. The cathode mounting and input leads must be designed for proper geometry at the cathode ends, heat dissipation, mechanical strength, and DC voltage breakdown strength.

The requirements placed on the magnetic circuit of a magnetron must be borne in mind throughout the design of the magnetron itself. Considerable effort may be expended in arranging for the magnet gap, and hence the required magnet, to be as small as possible. In "packaged" magnetrons the magnet pole pieces, which are built into the magnetron structure, must be designed to produce a field of proper configuration and to make the necessary external magnets feasible.

Generally, the design of the mechanical features of a device as complicated as the multiresonator magnetron oscillator is extremely important and must provide for structural strength under a variety of conditions, as well as cooling facilities and external protection of relatively fragile parts.

12. Reproduction of the British Magnetron

The problem undertaken at the Bell Telephone Laboratories immediately after the visit of the British delegation in October 1940 was the reproduction of the 10 cm. magnetron for study and for general radar use at the Bell Laboratories and the Radiation Laboratory at the Massachusetts Institute of Technology. The data available were contained in a drawing of a magnetron having six resonators and in an X-ray photograph of the magnetron used in the demonstration at the Whippany Laboratory, described in the INTRODUCTION. The X-ray photograph, reproduced in Fig. 45, showed a
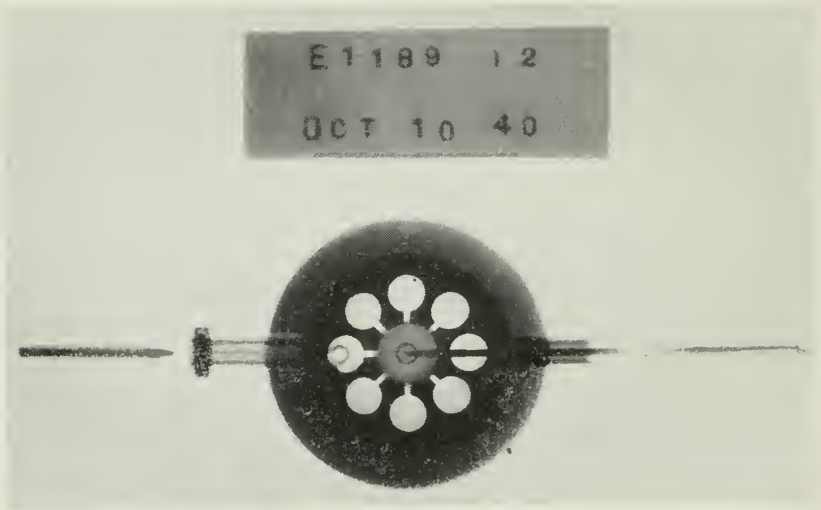
Fig. 45—An X-ray photograph of the 10 cm. magnetron oscillator brought to America by a British delegation in October 1940 and copied at the Bell Laboratories. It is the prototype of most magnetron oscillators in the centimeter wave region developed in Great Britain and the United States during the war.

resonator system having eight resonators. Since this arrangement was known to operate, it was adopted as the starting point for the work here.

In its first tests at Whippany, the British magnetron was pulse operated at about 10 kv. and 10 amps. peak current. The pulses were of 1 microsecond duration and recurred 1000 times per second. The magnetic field required was about 1100 gauss. The magnetron was loaded with a simple radiating antenna of unknown load impedance. Under these conditions the magnetron generated RF power estimated at the time to be greater than 10 kw.

The eight hole and slot type resonators of the British magnetron were spaced around an anode of 0.8 cm. radius. The resonator system, machined in a block of copper, was 2 cm. long. It was unstrapped, strapping

not being known at the time, and in its general features was much like that shown schematically in Fig. 1.

The output circuit of the British magnetron was also similar to that of Fig. 1. It had no particular transformer properties designed into it. The vacuum seal, made of copper, glass, and tungsten, was incorporated in the output coaxial line in very much the same manner as that shown in Figs. 60 and 61.

The cathode was a plain, oxide coated, nickel cylinder, 0.3 cm. in radius. It had nickel end disks of 0.5 cm. radius and was mounted on radial leads passing through glass vacuum seals like those shown in Fig. 61. The leads are placed diametrally across the resonator hole to minimize RF flux linkage to the cathode structure. Preliminary British results indicated that the cathode could be activated properly and would possess a reasonable lifetime under the original operating conditions.

The British magnetron had been designed for use with a magnet having a gap of about 1.75 in. and a pole face diameter of 1.25 in., producing a magnetic field of about 1500 gauss.

Several of the constructional features of the British magnetron were new. The cylindrical block of copper into which the resonator system was machined was used as the vacuum envelope. It was closed at either end by copper disk cover plates. The vacuum seal was made during the pumping and baking process by the alloying at the baking temperature of gold rings between the cover plate and block. The alloying was done at high pressure provided by a clamp bolted across the magnetron. Although no getter was used, satisfactory vacuum conditions could be maintained after seal-off.

By mid-November of 1940, a number of working reproductions of the British magnetron had been supplied in our Laboratories and to the Radiation Laboratory at M. I. T., and a program of study of the magnetron oscillator commenced. The work thus started was continued, on the one hand, to put the new magnetron into production, and on the other hand, to attempt to understand it, improve upon it, and extend its range of usefulness.

13. Magnetrons for Wavelengths of 20 to 45 Centimeters

13.1 *The 700A-D Magnetrons:* After the British 10 cm. magnetron had been successfully reproduced and an emergency program of research and development of multicavity magnetron oscillators commenced, the question immediately was asked: Can a multicavity magnetron be designed to operate near 40 cm. in the pulsed radar set under development in the Whippany radio laboratory? Clearly there now existed the possibility of much greater power than was possible with triodes at this wavelength with reasonable

life expectancy. The modulator of the radar set provided pulsed input power to the oscillator at about 12 kv. and up to 10 amps. peak current.[24]

The performance of the 10 cm. multicavity magnetrons appeared to make the development of such a generator at 40 cm. feasible. A straightforward enlargement of the 10 cm. magnetron by a factor of four was out of the question, however, as it resulted in a magnetron entirely too bulky, requiring a prohibitively large magnet. The development of the 700 mc/s magnetron oscillator thus involved departures from the British design. In particular it was found necessary to reduce the axial length of the resonator system to a considerably smaller fraction of a wavelength than in the 10 cm. design. The development involved design of the interaction space for maximum operating efficiency, the resonator system, for which both eight and six resonator structures were employed, and the output circuit for coupling into the existing radar system.

An early 700 mc/s multicavity magnetron design employed eight resonators of axial length less than one tenth wavelength; the 10 cm. design was about one fifth wavelength long. Operating models initially produced approximately 10 kw. of RF power near the desired frequency. It was found, however, that a smaller and lighter magnetron could be made to operate at the same voltage if the number of resonators were reduced from eight to six, permitting smaller anode and cathode radii [equation (16) in PART I]. The weight and over-all diameter was further reduced by use of elongated holes in the hole and slot resonators. This change resulted in the resonator system used in the 700A-D magnetrons (see Figs. 41 and 46). Each hole is made by boring two intersecting cylinders in the resonator block as may be seen in Fig. 46. No difficulty was encountered in achieving the desired frequency. The frequency differences between the four coded magnetrons near 700 mc/s were achieved by variation of the resonator slot width.

The separation of mode frequency between the $n = 3$ mode (π mode) and the nearest other mode is of the order of 3 per cent. Although this is small compared to that obtainable in strapped magnetrons, it is greater than that for the early unstrapped magnetrons near 10 cm. This is reflected in greater operating efficiency.

The cathode in the 700A-D magnetrons is supported, as in the British magnetron, by radial leads extending across the center of one of the hole and slot resonators. The cathode diameter was varied in an experiment designed to determine the value for maximum operating efficiency. Early experiments of this type involving measurements of output power and efficiency were quite crude and conclusions from their results were by no means as significant as those based on measurements of frequency. The primary

[24] This radar development is discussed in: W. C. Tinus and W. H. C. Higgins, "Early Fire Control Radar for Naval Vessels," *Bell Syst. Tech. Jour.*, 25, 1 (1946).

difficulty lay not in the actual measurement of power or voltage but in the fact that the magnetrons were not loaded in a reproducible fashion. It was considerably later that load impedance measurements were made and used in evaluating magnetron performance. In many early studies the effect of load on operation was not sufficiently disentangled from the effects of other things. In spite of these inadequacies, however, it was generally possible to distinguish a good design change from a bad one, and much of value was gained in early work.

The cathode diameter used in the 700A-D magnetrons is given in TABLE I along with other data on these and other magnetrons in the 20 to 45 cm.
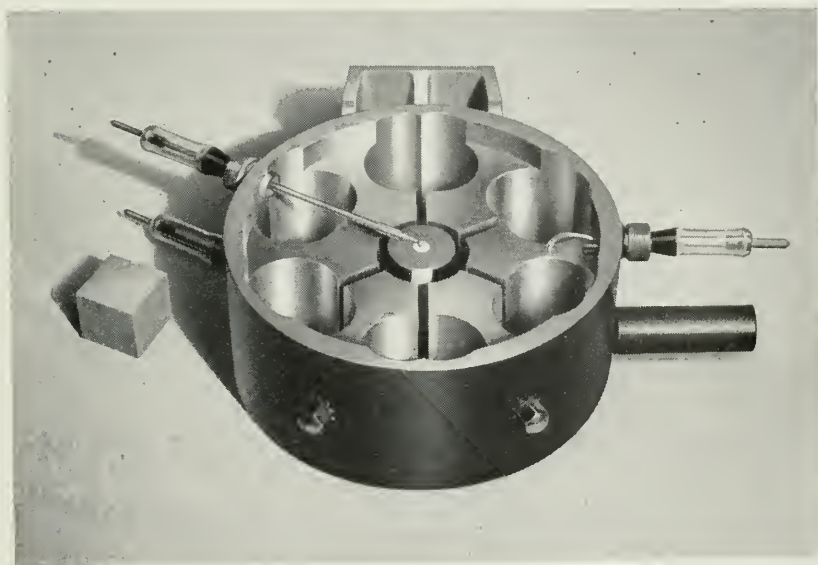
Fig. 46—An internal view of a 700A-D magnetron (40 kw., ∼700 mc/s) showing the unstrapped resonator system of six hole and slot circuits, the cathode, the cathode end disks and support leads, and the output coupling loop and lead.

wavelength range. It should be noted that the optimized ratio r_c/r_a is 0.300 as compared to 0.375 in the British magnetron having eight resonators. Plain oxide coating is used on the cathode. Life expectancy is thousands of hours.

As may be seen in Fig. 46, the output coupling is accomplished by means of a loop in the end space of the structure. The loop is connected to an anode segment. It is driven by direct coupling to the anode segment and by coupling to the magnetic flux linking the two adjacent resonators. The output circuit was not designed to operate into a matched output line, however, and external impedance transformation must be incorporated into the load line.

TABLE I

MAGNETRONS FOR WAVELENGTHS OF 20 TO 45 CENTIMETERS

	700A-D Unpackaged	728A-J Unpackaged	5J23 Unpackaged	4J21-25 Unpackaged	4J26-30 Unpackaged	4J42 Unpackaged Tunable	4J51 Unpackaged Tunable	5J26 Unpackaged Tunable
N	6	8	8	8	8	6	8	8
r_c (in.)	0.160	0.266	0.266	0.218	0.230	0.199	0.266	0.375
r_a (in.)	0.689	0.687	0.709	0.582	0.612	0.689	0.687	0.687
h (in.)	1.576	1.716	2.360	1.940	2.040	1.451	1.500	1.940
Magnet gap (in.)	2.980	3.290	3.990	3.540	3.640	2.983	3.290	3.640
Weight (lb.)	12.5	13.0	16.5	15.0	15.0	16.5	14.5	18.5
Resonators	hole and slot	hole and slot	hole and slot	hole and slot	hole and slot	hole and slot	hole and slot	slot
Unstrapped λ (cm.)	43.0	~26.0	~21.5	~18.0	~19.0	~38.0	~23.5	10.3
Straps	none	double ring	echelon wire	double ring	double ring	wire	double ring	double channel
λ (cm.)	43.0	32.1	28.6	22.8	24.0	43.0	32.1	23.4
f (mc/s.)	720–680	970–900	1056–1044	1350–1280	1280–1220	670 to 730	900 to 970	1220 to 1350
Nearest mode	$n = 2$	$n = 3$	$n = 3$	$n = 3$	$n = 3$	$n = 1$	$n = 1$	$n = 3$
λ separation (%)	−3	~−30	~−20	−20	−20	−16	+4	~−60
Tuning	—	—	—	—	—	resonator capacitance	resonator capacitance	strap capacitance
Δλ (%)	—	—	—	—	—	10.2	7.5	10.3
Tuner travel (in.)	—	—	—	—	—	0.100	0.080	0.154
Q_0	>5000	~4500	~3200	2800	2800	1600→2500	3500→4500	700→1800
Q_{ext}	~280	170	150	170	180	285	215	210
η_c (%)	~95	~96	~95	94	94	87	95	82
Output circuit	coaxial	coaxial	coaxial	coaxial	coaxial	coaxial	coaxial	coaxial
V (kv.)	12	19.0 21.0 24.5	16.5 19.0 24.5	15.5 22.0 26.5	16.5 23.0 27.0	12.0	23.0	27 27
I (amps.)	10	19 20 28	20 24 33	25 40 48	25 40 46	9	20	46 46
B (gauss)	650	1000 1100 1200	800 900 1100	900 1200 1400	900 1200 1400	650	1100	1400 1400
τ (μs.)	2	1 1 1	1 1.5 1	1.5 1 5	1.5 1 5	1.5	1	1 5
pps.	1000	1000 1000 1000	2000 1000 1000	1000 1000 200	1000 1000 200	2000	1000	1000 200
P_o (kw.)	40	210 260 400	170 250 475	175 440 640	175 470 700	30	285	600 600
η (%)	33	58 62 58	55 55 59	48 50 50	48 51 56	28	62	48 48
η_e (%)	~35	~61 ~65 ~61	~58 ~58 ~62	51 53 53	51 54 60	32	65	58 58
PF (mc/s.)	~1.2	2.5 2.5 2.5	3.0 3.0 3.0	3.4 3.4 3.4	3.0 3.0 3.0	1.2	1.9	3.0 3.0

In mechanical construction the 700A-D magnetrons involved techniques like those described above. The input and output leads included copper to glass to tungsten seals much like those in the reproductions of the British magnetron. The end covers were sealed to the resonator body by means of the gold ring technique employed in the British magnetron.

The 700A-D magnetrons are limited in frequency to the four 10 mc/s bands between 680 and 720 mc/s, respectively. These magnetrons operate at 12 kv. and 8 amps. peak current input at a magnetic field of 650 gauss. Over-all efficiency ranges between 30 and 40%, which is better, as has been explained, than that attained with unstrapped 10 cm. magnetrons. Other data of interest are given in TABLE I.

One feature which is immediately apparent from the rated operating conditions of the 700A-D magnetrons is the fact that the ratings are not nearly as high as one might expect from the size of the magnetron. Back bombardment of the cathode at considerably greater input power could easily be handled. The difficulty lay in the fact that it was impossible to drive the magnetrons in the π mode to much greater currents than the rated currents. If the attempt is made to drive the magnetron harder it either refuses to oscillate at all or oscillates in another mode. This phenomenon has been the single deterrent in the development of higher power magnetrons at wavelengths greater than 20 cm. It is now recognized as a starting time phenomenon having to do with the rate at which oscillation builds up and the rate at which pulse voltage is applied (see Section 10.6 *Oscillation Buildup—Starting*). What has been done in studying the phenomenon and in magnetron design to circumvent it will be discussed in some detail in connection with the 5J26, the tunable replacement for the 4J21-30 series.

In quantity production the 700A-D magnetrons presented new problems, all of which arose because of its size. The oxide coated cathode, having a relatively large surface area, gave off a considerable quantity of gas during cathode activation. In as much as the massive copper anode could be outgassed only by a long baking process at temperatures below the softening point of the glass parts, difficulty with magnetrons "going soft" after seal-off was encountered initially.

The development of the 700A-D magnetrons was carried on simultaneously with early studies at 10 cm. and with the early attempts to produce power at 3 cm. A number of auxiliary experiments were undertaken which, although they were not a part of the specific magnetron development, contributed results of considerable value complementary to those obtained at the shorter wavelengths. In particular, these experiments had to do with the technique of measurement and of magnetron scaling.

Before the invention of straps the 700A-D magnetrons were scaled to 10 cm. to explore the possibilities of a more efficient magnetron design at this wavelength. Straps were introduced before the completion of the experiment. The resultant strapped magnetron having six resonators was very efficient—60 per cent—but required a high magnetic field as can be seen by referring to equation (16) of PART I. Like other magnetrons, the 700A-D became much more efficient when strapped. At the normal test point the efficiency ranged around 50 per cent, while at higher magnetic field and voltage, 75 per cent over-all efficiency was achieved. The introduction of straps into the manufactured design was not undertaken.

One further experiment of interest arose during the development of the 700A-D magnetrons from the desire to measure the gas pressure in a sealed-off magnetron. The non-oscillating magnetron itself was used as an ionization manometer. With the magnetic field set at a high value above cutoff and under conditions of no RF oscillation, electrons which arrive at the anode can do so only after having lost energy by collision with a gas molecule. Under these conditions the anode current is directly proportional to the pressure.

Although by present standards the 700A-D magnetrons might appear somewhat crude and inadequate, they nevertheless have an important place in the story of wartime magnetron development. They filled an immediate need in the radar system for which they were designed, providing the U. S. Navy with a radar set which saw service in a number of crucial engagements. Furthermore, the development of the 700A-D magnetrons provided invaluable experience.

13.2 *The 728A-J Magnetrons:* The 728A-J magnetrons were developed for fire control and search radar systems to supersede those which had used the 700A-D magnetrons In these n w systems a magnetron generator was to be required which could deliver 200 kw. peak output power in the frequency range 920-970 mc/s (later extended to 900 mc/s).

In an early design, the resonator system had eight resonators and was strapped with wire straps in the early British configuration [see Fig. 24(a) of PART I]. The anode length was 4 cm., the same as was used in the 700 A-D, which on a wavelength basis was about 2/3 that used in the British 10 cm. magnetron. The first models were designed for operation at pulse voltages of about 27 kv. When, subsequently, it was decided to reduce this voltage, a redesign involving a reduction of size of the interaction space became necessary. Since more had been learned about the technique of strapping in the meantime, it was decided that straps of the double ring type [see Fig. 24(d) of PART I] should be used in the new design. At first, the straps were set on the ends of the anode structure projecting into the end spaces but were later recessed into channels cut into the copper resonator

structure for the purpose of electrostatic shielding from electrons in the interaction region. The frequency range required was spanned by the use of anode structures having three different slot widths for the primary frequency separation, small additional frequency shifts being obtained by small distortions of the straps. Resonant frequencies of magnetron resonator systems were now being determined prior to sealing for pumping by measurements like those described in PART I, during which any necessary strap adjustment could be made.

The cathode was a plain, oxide coated, nickel cylinder much like that used in the 700A-D magnetrons. The heater inductance was considerably higher than that of any previous cathode assembly. It was found that sudden and severe transient conditions, such as those imposed by a momentary internal arc between cathode and anode, would cause relatively high voltages to develop between the cathode and the open end of the heater. This could break down the heater insulation and cause either open or short circuits. This difficulty was minimized by incorporating in the driving equipment a condenser across the heater and an RF choke in series with the heater. Before final design specifications were submitted, the input leads and cathode structure were completely redesigned to provide greater rigidity and strength. To withstand violent shock and vibration, the structure was designed to have as high frequencies of mechanical resonance as possible. The structure looked much like that to be seen in the 5J23 magnetron of Fig. 49. Direct mechanical injury to the input leads is prevented by the use of a heavy glass housing.

The output circuit in early experimental models was a coaxial type fed by a loop in one of the resonators. The central conductor was a tungsten rod to which the glass seal was made and to which the inner conductor of the load coaxial was clamped. When the resonator system was redesigned for lower voltage, a new design of output circuit was made in which was used a choke or contact-free load coupling like that designed for the 720A-E. This removed the possibility of stress being applied to the glass of the output seal at either the inner or outer conductors. Except for the critical dimensions determined by frequency, the output circuit is identical to that used on the 5J23 shown in the photograph of the cut-away model of Fig. 49.

In Fig. 47 is shown a schematic diagram of this type of coupling. On both inner and outer conductors an electrical short circuit is produced at the gap between magnetron and load coaxials by folded low impedance coaxial sections incorporated into the bodies of the conductors. In the outer conductor a half wavelength section folded once is employed. In the section shown at (a), the joint is made at the current node in the choke section by the outer of two cylinders which project from the load end of the coupler

into the magnetron lead. In the partial section shown at (b), this outer cylinder is not used and the joint occurs at the end of the section where there is a current antinode. In this method there exists the greater possibility of sparking should the coupling not be clamped tightly. A folded quarter wavelength section is built into the inner conductor. If the wavelength is short enough, as in the 5J26 and the 720A-E (see Figs. 58 and 63), this section need not be folded, the inner post on the magnetron center conductor can be eliminated and replaced by a solid center conductor on the load side.
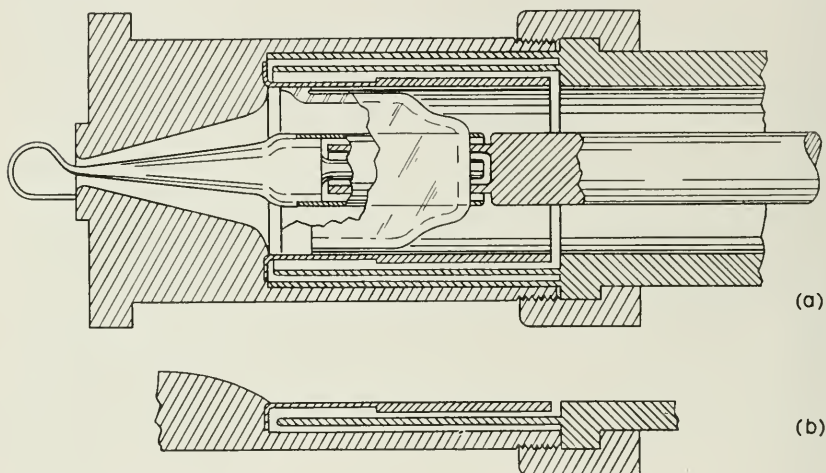
Fig. 47—A schematic diagram of the type of coaxial output circuit used in a number of magnetrons of wavelength 10 cm. or greater. Of particular interest are the means of contact-free or "choke" coupling employed in the inner and outer conductors, consisting of a folded concentric line section which presents zero impedance at the gap in the conductor. (a) and (b) represent two variations of the "choke" in the outer conductor as explained in the text.

The more recent designs of magnetrons for wavelengths greater than 10 cm. have some form of this coupling.

The impedance required at the output coupling to load the 728A-J magnetrons for sufficient power output necessitated a rather high standing wave in the output line. This standing wave was provided by a transformer built into the radar system to which the magnetron was attached. This caused no trouble since the output power was below the point where RF voltage breakdown in air in the line or coupling would occur. The press of time necessitated the adoption of this output circuit although "preplumbing" by incorporation inside the vacuum of the necessary transformer action for coupling directly to a matched load line would have been preferable. Such a design was executed but its completion came too late for its incorporation into the manufacturing specifications.

Ten different magnetrons, coded the 728A-J, were put into manufacture. These covered the frequency range from 900 to 970 mc/s, a 7 mc/s range being allotted to each code type. The operating characteristics of the final design together with other pertinent data are tabulated in TABLE I. An external view of the magnetron is shown in Fig. 48. A maximum current limitation for satisfactory operation in the π mode was also encountered in this magnetron but at current values above the 20 ampere operating point
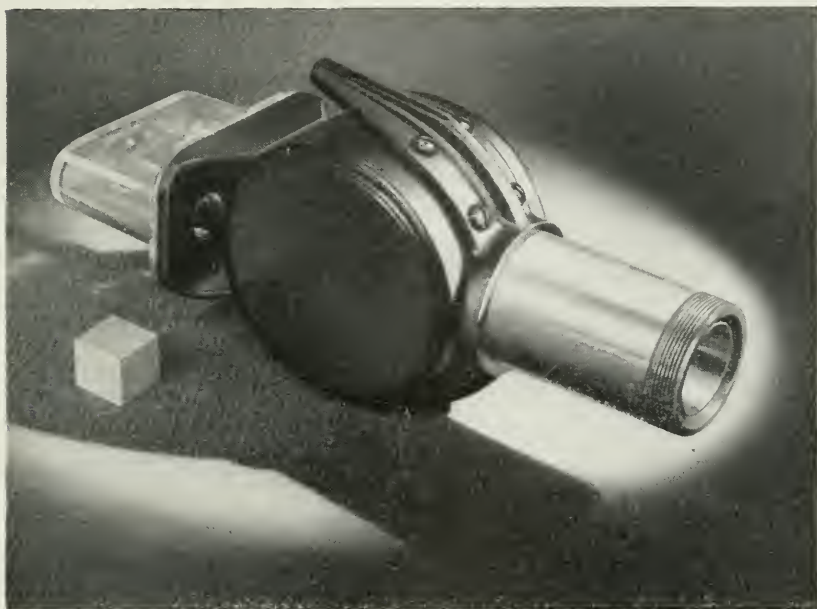
Fig. 48—An external view of a 728A-J magnetron (275 kw., ∼930 mc/s). The concentric cylindrical sleeve to be seen inside the output circuit coupling, from which the magnetron is supported, is a part of the "choke" (see Fig. 47). Note the heavy glass protective housing over the input leads.

at which the required output power was attained. Because of the shorter wavelength, the current limitation was not as severe as that encountered in the 700A-D.

The 728A-J magnetrons were driven by a spark wheel, line type modulator. The greater tendency of the driving voltage to overshoot with this modulator, by virtue of the slowness of buildup of RF oscillation, was reduced by the use of a series resistance and capacitance network coupled across the magnetron input like that used earlier with 720A-E magnetrons at 10 cm.

13.3 *The 5J23 Magnetron:* The request for the Laboratories to develop

a semiportable radar pack set and a searchlight control radar involved the design of a new magnetron in the frequency range 1050 to 1110 mc/s to operate at a maximum pulse voltage input of 25 kv. with power output of at least 200 kw. Since these requirements, except for frequency, were essentially those of the 728A-J magnetrons already under development, the new work closely paralleled that already in progress.

Because of the maximum current limitation being encountered in the development of longer wavelength magnetrons, a relatively longer anode was designed for the new magnetron, coded the 5J23, thus insuring satisfactory operation at higher current and input power. The resonator system was "heavily" strapped with echelon wire straps [see Fig. 24(c) in PART I]. Coarse frequency variation was accomplished by the use of straps at different heights above the surface of the anode block. Finer frequency adjustments were made in "pretuning" the structure, before sealing, by small displacements of the wire straps already in position.

Output circuit work on the 5J23 was done together with that on the 728 A-J magnetrons and went through the same series of developments. The output circuit used in the manufacturing design is identical with that of the 728A-J except for the dimensions critical to wavelength.

For operational reasons, only one magnetron, the 5J23, in the frequency range 1074 to 1086 mc/s, was coded. Fig. 49 shows a cut-away view of its internal structure. Performance characteristics and other data are to be found in TABLE I. Note that the 5J23, by virtue of its greater anode length and higher frequency, has a higher critical current, I_c, above which π mode operation fails, than do the 700A-D or 728A-J magnetrons.

13.4 *The 4J21-30 Magnetrons:* When development work on the 5J23 had just been completed, an international agreement on frequency allocation made the frequency range of the 5J23 unavailable for radar purposes at some future date. Consequently the work was to be redone for operation in the band 1220 to 1350 mc/s. The 5J23 and the equipment in which it was used were to continue in manufacture and be used in the period necessary for the development of the new equipment and magnetrons. In addition to the frequency change, power output demands on the magnetron were raised considerably, it being required that the new magnetrons produce at least 500 kw. output power. Although the maximum current limitation had seriously affected maximum output power capabilities of the magnetrons in this wavelength range, it appeared that the new high power demands were within reach by virtue of the higher frequency at which the new magnetrons were to operate.

The new assignment came just upon completion of the 5J23 and made it necessary to build or acquire new laboratory equipment, CW oscillators,

wavemeters, output couplings, and loads. On the other hand, the new development did provide the opportunity to make use of new ideas and developments, some of which, although already completed, could not be used in magnetrons then in manufacture.

When work on this new series of magnetrons was commenced, the art of magnetron development had reached the stage where for the first time it was possible to make a reasonably satisfactory approach to the problem of a new design. The principles of magnetron scaling had been worked
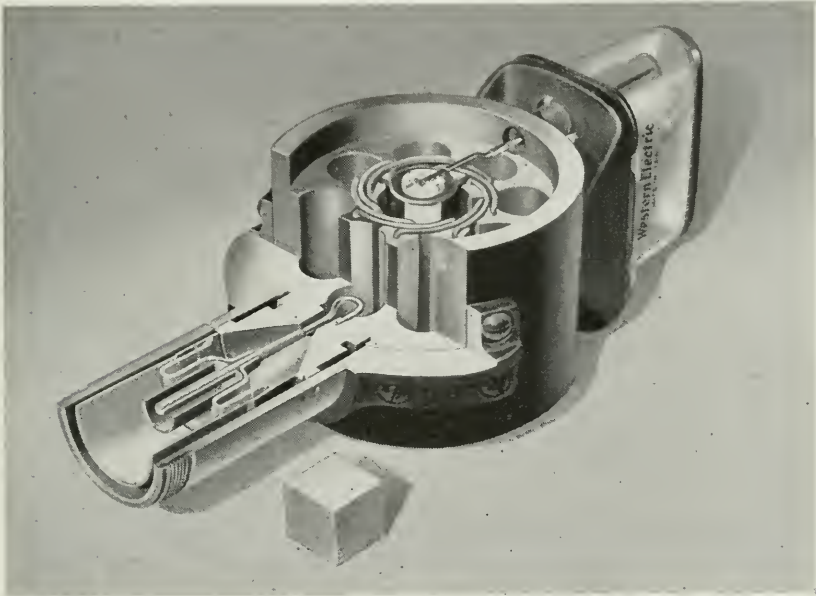
Fig. 49—A sectioned 5J23 magnetron (275 kw., ∼1050 mc/s) showing, among other things, the echelon type of wire strapping and coaxial output circuit with contact-free load connector.

out and tested in detail, enabling one to make use of results obtained or verified at different frequencies. The mode frequency spectrum of a magnetron resonator system was now understood and the means of controlling it by strapping were in use. The problem of the output circuit was in hand. Magnetrons were being "preplumbed" and the problems of reproducibility were being solved. Furthermore, the method of studying each of these magnetron characteristics by CW measurements on resonator blocks and output circuit models had advanced to the point where it was possible to complete a design in detail before an operating magnetron had been constructed. Such a procedure was adopted for the new series of

magnetrons, coded the 4J21-30, in the band 1220-1350 mc/s. A satis-
factory resonator system giving proper operating voltage, π mode frequency,
and mode frequency separation was designed prior to and incorporated in
the first operating model. Similarly, a satisfactory output circuit for proper
loading over the broad band of frequencies and the design of mechanical
features involved in the cathode, cathode mount, cooling facilities, output
circuit, straps, and resonator system were provided.
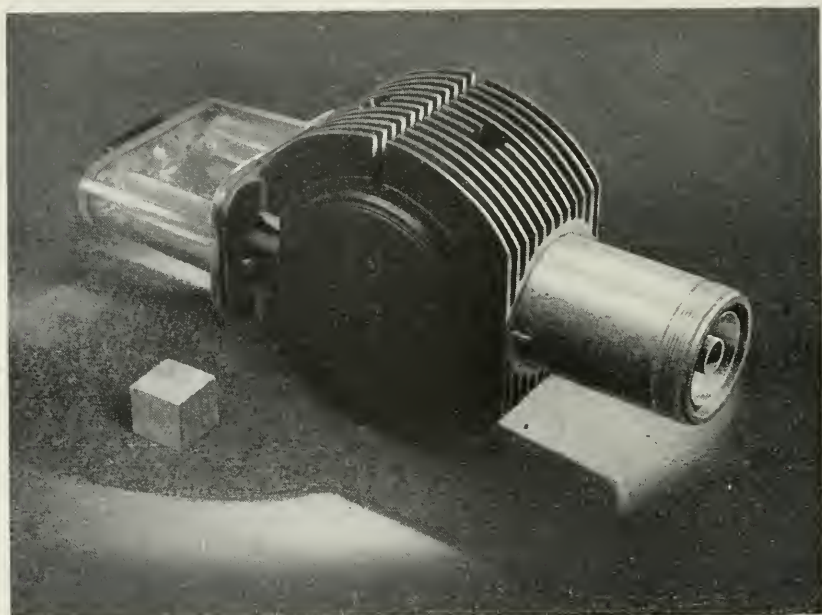
Fig. 50—An external view of the 4J28 magnetron (one of the 4J21-30 series, 600 kw.,
~1280 mc/s). Note the cooling fins attached to the magnetron body.

The new interaction space and resonator system were scaled from a com-
promise between the 728A-J and the 5J23 designs. Shielded, double ring
straps like those of the 728A-J were adopted. Two anode structures
differing somewhat in resonator dimensions and size of interaction space
were used, one for the short wavelength half of the band (4J21-25), the other
for the long wavelength half of the band (4J26-30). The different frequen-
cies within each of these groups were attained by the use of straps of different
widths.

A single design of output circuit for the entire frequency range necessitated
making the transformer properties as independent of frequency, or as broad
band, as possible. The output circuit has three parts. At the outside of the

magnetron is the choke or contact free coupling to the load coaxial. The general type of design used in previous magnetrons was followed here. The transformation from the load coaxial to the coaxial which breaks into the resonator at the coupling loop was made by use of a tapered line in which the characteristic impedance was kept constant. The proper loading was accomplished by adjustment of the loop size. It was found possible to achieve a sufficiently low loaded Q in this way within the bounds of a satisfactory geometrical arrangement. Each of these design procedures was
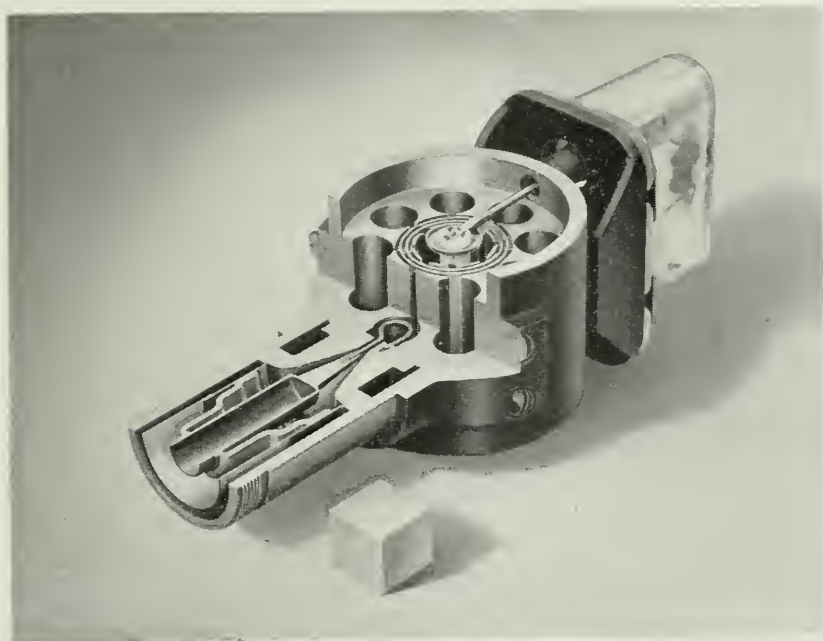
Fig. 51—An internal view of a 4J21-30 magnetron (600 kw., $\sim$1280 mc/s). Of special interest are the recessed double ring straps and the output circuit in which the characteristic impedance of the coaxial line from loop to load is approximately constant.

carried out by CW measurement of impedance, looking into a model through the output circuit. Special models of the tapered line and choke coupler were also studied.

When operating magnetrons were constructed, except for minor changes, no redesign was necessary. The first working models performed satisfactorily at the required pulse voltage, current, magnetic field, power output, and pulling figure. It was possible to get more than 750 kw. output at better than 50 per cent efficiency from these magnetrons.

An external view of the final design may be seen in Fig. 50 and a photograph of a cut-away model in Fig. 51. Operational and other data are given

in Table I. A typical performance chart of one of the series is shown in Fig. 52. The maximum current boundary limits the output to values well below those at which other factors, such as cathode dissipation, become restrictive.
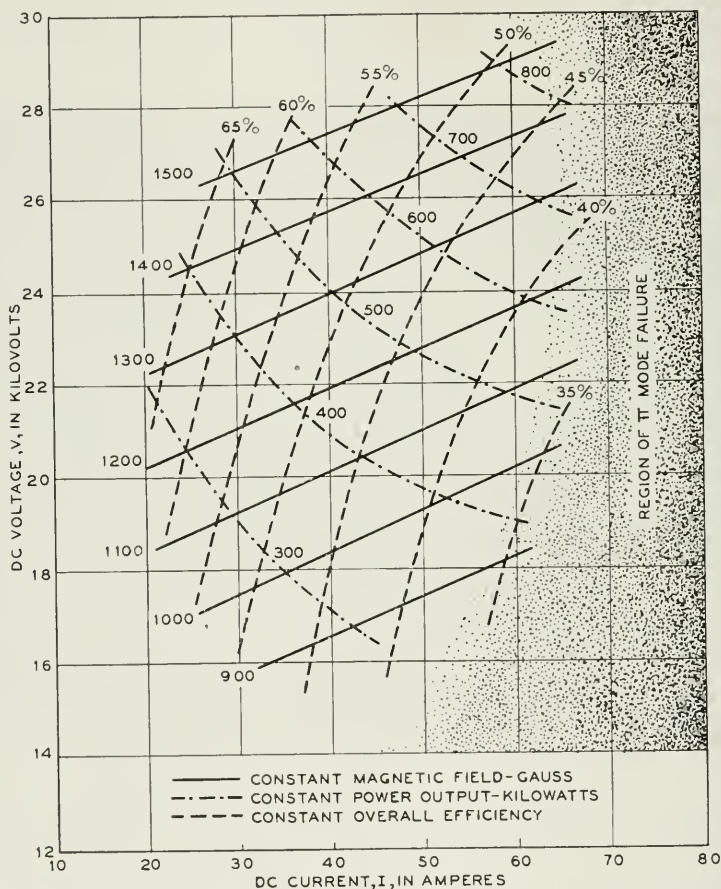
Fig. 52—A typical performance chart of a magnetron of the 4J21-25 series (∼1315 mc/s).

During operation in the high power, spark gap, line type modulator with which the magnetron was to be used, it was found that the pulse voltage applied to the cathode input leads was sometimes sufficient to cause flashover. If the arc persisted, vacuum failure from cracking or puncturing of the glass resulted. The external flash-over was accompanied by internal arcs between cathode and anode which were injurious to the life of the

magnetron. To circumvent this difficulty, the input leads and their glass housing were redesigned. The glass of the leads was lengthened and folded to provide a greater surface distance over which the arc must strike. Although the redesign was limited somewhat by the space available in the transmitter unit, it was possible to make leads which could stand 60 cycle peak voltages of almost twice the normal pulse voltage applied to the magnetron. This step all but eliminated flash-over in operation. To take care of the occasional flash-over, a spark gap across which the breakdown could occur was added in the equipment.

The 4J21-30 magnetrons are illustrative of carefully designed nontunable magnetrons of wavelengths 20 to 30 cm. developed at the Bell Laboratories. They were the first magnetrons designed here completely on the basis of CW impedance measurements.

14. Tunable Magnetrons for Wavelengths of 20 to 45 Centimeters

14.1 *General:* As fixed frequency magnetrons became available in the various frequency bands designated for radar use, the interest quite naturally turned to the development of tunable magnetrons. The operational reasons for this were to enable one to set the frequency of a radar system at will, thus avoiding interference between sets in a large group of aircraft on naval vessels, to enable one to vary the frequency at will from time to time or as the need arises to avoid jamming, and to permit the stocking of fewer magnetrons to cover a given frequency band.

Early work on tunable magnetrons at the Bell Laboratories was done with 10 cm. models. Although no such developments were carried to the stage of production, the ideas and techniques evolved were used at other frequencies. Somewhat later the Bell Laboratories committed itself to a program of development of tunable magnetrons for pulsed radar use in the 20 to 45 cm. wavelength range.

The program initially was not directed toward the goal of some particular magnetron of fixed specifications. Rather, it was the intention to explore the field of possible tuning methods and to find that one which appeared both electrically and mechanically to be best suited for large magnetrons. Work in this initial stage was done on anodes of the 4J21-30 series and may be divided into two main channels characterized by the degree of symmetry involved in the tuning scheme. In one, the tuning of the entire anode block is accomplished by modification of a single one of the resonators and may thus be characterized as unsymmetrical tuning. In the other, each resonator of the multiresonator block is tuned in a symmetrical fashion.

Among the unsymmetrical types studied were those employing an auxiliary loop in one resonator connected to a reactive element which could be either a

coaxial line terminated in a plunger or a variable capacitance, or a coaxial to wave guide junction with movable plungers in the wave guide. Another form involved the deformation of one of the resonators itself to accommodate a prism shaped tuning element which was moved from outside the vacuum through a diaphragm arrangement.

Although considerable effort was expended in the study of unsymmetrical tuning schemes and much learned about them in the course of this work, they have two rather fundamental drawbacks which caused them to be replaced by the symmetrical types in which these defects could be eliminated. In the first place, it is difficult by this means to obtain the desired range of frequency change. Secondly, by virtue of asymmetry, operation in other modes is difficult to avoid except over rather narrow frequency ranges. Those types in which part of the tuning circuit is outside the vacuum envelope have the additional drawback of bringing high RF voltages into structures in air where very high standing waves are necessary and breakdown difficult to avoid.

It was found possible to circumvent each of these difficulties by using symmetrical tuners. Some such tuners were like those tried on 10 cm. models and involved spider-like straps connecting alternate segments to two common points on the axis of the tube. These points in turn were connected to the center and outer conductors of a coaxial line. After passing through a vacuum seal, this coaxial line connected to a reactor like those used in the unsymmetrical schemes.

In the general arrangement finally used, a tuning member was mounted in the end space. The structure included a vacuum diaphragm mounted in the end cover of the magnetron. The tuning was effected by variation of the capacitance between the resonator system and the tuning member.

In unstrapped magnetrons, the tuning was accomplished by moving a tuning member of the general shape of a "cookie cutter" in a single groove cut into the slotted portion of the anode structure. Later, this range was increased by using two slots and a tuning member having two concentric rings. This type of tuner was used on both the 4J42 and 4J51 magnetrons to be discussed. It is the first of the two types of tuning by variation of capacitance discussed in PART I of this paper. Considerable effort was expended to extend the frequency range of tuning by strapping the untuned end of the resonator system and by variations of the resonator dimensions on which relative mode frequencies depend. It was found possible to obtain a tuning range of better than ±5 per cent, but it appears quite difficult to exceed this amount appreciably in high voltage magnetrons using this tuning means.

In the strap tuning scheme, use was made of straps of channel or U-shaped cross section, inside which the double rimmed "cookie cutter" element is moved. The tuning arrangement of the 5J26 magnetron is of this type. It is the second of the two types of tuning by variation of capacitance discussed in PART I.

Quite early in this program of development it became apparent that the armed services would require all these tunable magnetrons to be mechanical and electrical replacements for existing fixed frequency magnetrons, interchangeable in all respects. This requirement, in addition to the facts of electrical design, indicated that the "cookie cutter" type of tuner represented the line of attack most likely to succeed. The capacitance in the anode structure or in the straps is the frequency determining circuit parameter most easily varied, as it is the most accessible in these large size magnetrons. The tuning element may be moved conveniently by a mechanism, compact in the axial dimension, mounted in the end space and cover of the magnetron. These are factors of prime importance in "unpackaged" magnetrons where over-all axial length is important. Magnetrons in this wavelength range are sufficiently large to make such a mechanical design feasible.

The magnetrons of wavelength 20 to 45 cm. for which tunable replacements were designed were: the 700A-D series (680-720 mc/s), the 728A-J series (900–970 mc/s), and the 4J21-30 series (1220-1350 mc/s). Although it was preferable to span the 4J21-30 band with one tunable replacement, the use of two tunable replacements, each covering half the band, would have been satisfactory had it been necessary. In the development of each of the tunable replacements, improvements and modernizations which could be incorporated within the limitations of mechanical and electrical interchangeability were made.

14.2 *The 4J42 Magnetron:* The tunable replacement for the 700A-D series is the 4J42. It covers an increased frequency band of 660 to 730 mc/s. The tuning is accomplished by variation of the intersegment capacitance of the anode structure itself, the first of the two capacitance tuning methods listed above. Frequency varies nearly linearly with displacement of the tuning member. The tuning range, although adequate for the purposes of the 4J42, is limited by interference of other modes as was discussed in PART I. In this respect it is not a great deal better than can be attained by unsymmetrical means. The advantages of symmetrical capacitance tuning enumerated above made it the logical choice nevertheless. A satisfactory tuning range of the π mode without interference from other modes was achieved by strapping the end of the magnetron opposite that at which the tuning mechanism is mounted. This necessitated changes in the resonator

dimensions of the 700A-D magnetrons to obtain the frequency range desired. The cathode structure was improved, making use of later developments. Its diameter was changed from 0.8 cm. to 1.0 cm. in order to increase the maximum current to which operation in the π mode is possible. The anode diameter was retained; the operating conditions were not appreciably affected. It was necessary to retain the original form of output circuit.
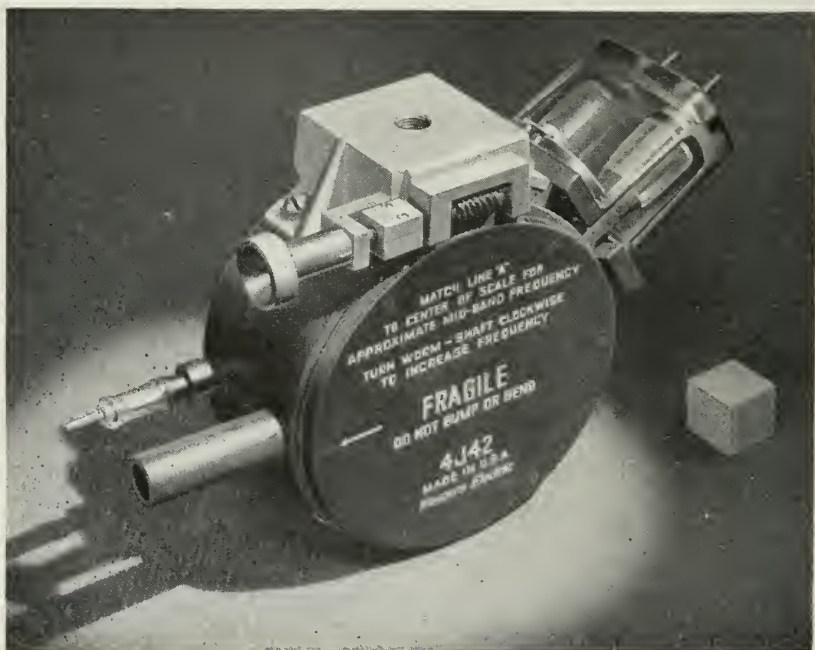
Fig. 53—The 4J42 tunable magnetron (40 kw., 660 to 730 mc/s)—the tunable replacement for the 700A-D magnetrons. The worm drive in the mounting bracket operates the tuning mechanism enclosed in the end cover of the magnetron. Note the improved input leads and protective housing (see Fig. 46).

In Fig. 53 is shown an external view of the 4J42. In Table I are given operational data and other characteristics. The performance of the 4J42 is as good or better than that of the corresponding magnetron in the 700A-D series.

14.3 *The 4J51 Magnetron:* The 728A-J series of fixed frequency magnetrons was replaced by the 4J51, covering the frequency band of 900 to 970 mc/s. An external view of the magnetron is shown in Fig. 54. As in the 4J42, the resonator system is a modification of that used in the fixed frequency prototype, strapped on the untuned end by double ring strapping. Satisfactory mode frequency separation is maintained by increasing the

strap capacitance of the single pair of rings to approximately twice that of one pair in the 728A-J.   The tuning scheme is the same as that used in the 4J42.   It may be seen in the photograph of a cut-away model in Fig. 55.

The output circuit of the 4J51 incorporated the broad band transformer properties developed for the output circuit of the 4J21-30 series.   This necessitated a larger loop size in the 4J51 than was used in the 728A-J.
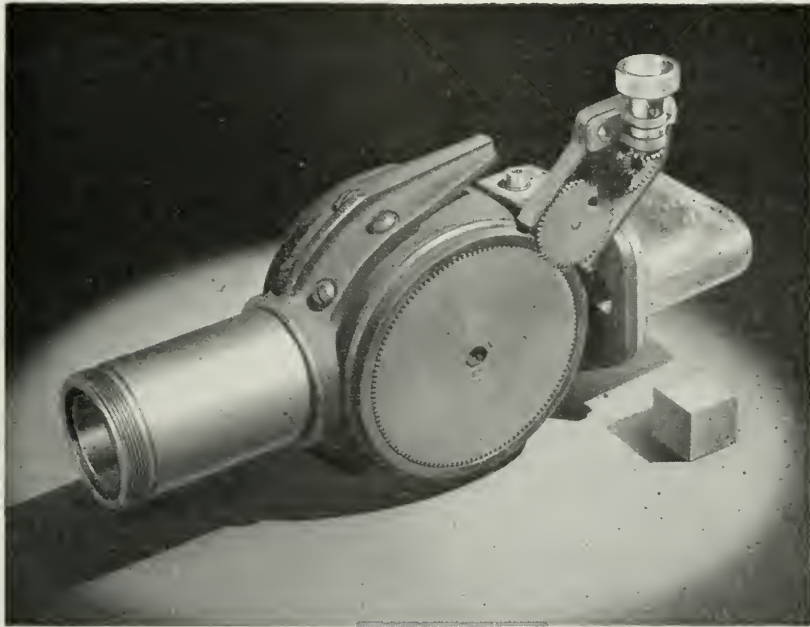
Fig. 54—The 4J51 tunable magnetron (275 kw., 900 to 970 mc/s)—the tunable replacement for the 728A-J magnetrons.

These features may be seen in Fig. 55.   The output circuit was designed for coupling to a matched load line.

TABLE I includes relevant data on the 4J51 magnetron.   In Fig. 56 are shown power output and mode frequency tuning curves for the magnetron. Fig. 56 may be taken as representative of the type of tunable magnetron in which the intersegment capacitance of the resonator system is varied. The frequency range is limited by the crossing of the π mode frequency curve by those of other modes.   The power output is reduced appreciably at the crossing point of the $n = 3$ mode component.   Throughout its tuning range, indicated in Fig. 56, the 4J51 magnetron operates as well or better than its fixed frequency predecessors.

14.4 *The 5J26 Magnetron:* Replacement of the 4J21-30 series of mag-

netrons with a tunable equivalent necessitated a tuning range which was found to be unattainable with the scheme used in the 4J42 and 4J51 magnetrons. Two tunable replacements employing the "tuned segment" type of tuning, each spanning half the band, were constructed as "insurance" should the efforts to develop a single replacement having the required range fail. However, it was found that by tuning the straps as has been de-
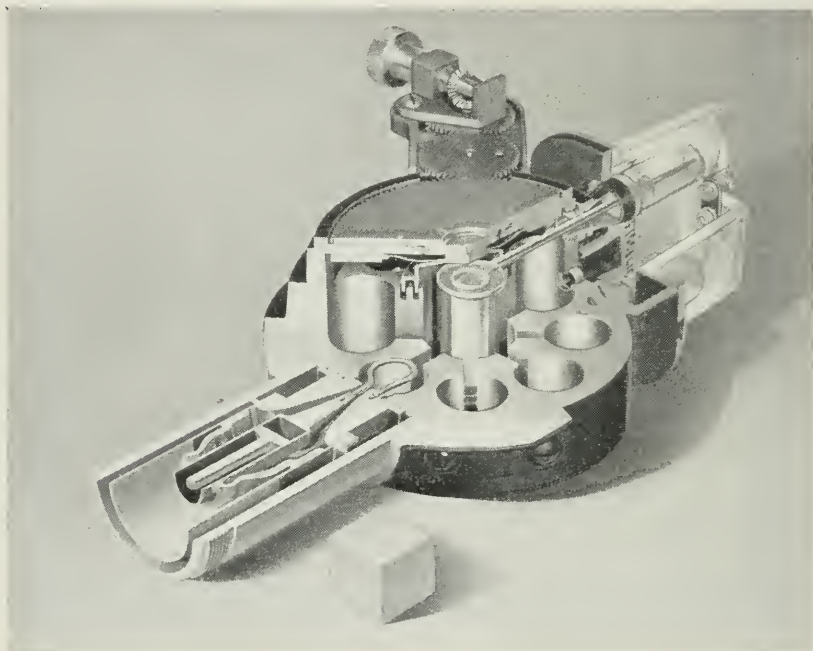
Fig. 55—A view of a cut-away 4J51 tunable magnetron (275 kw., 900 to 970 mc/s). Note the tuning member consisting of two concentric rings which are moved up and down in the grooves in the segments of the anode structure. The details of the tuning drive mechanism, including the flexible vacuum diaphragm, the axial screw, the nut and ball bearing, and the gear drive are to be seen. The resonator system is strapped on the end not seen in the figure.

scribed in PART I a single tunable magnetron for the entire 4J21-30 series covering the band from 1220 to 1350 mc/s could be provided. An external view of the resulting magnetron, coded the 5J26, is shown in Fig. 57 and an internal view in Fig. 58.

As can be seen in Fig. 58, the 5J26 magnetron includes radical departures in design in addition to the tuning scheme. Slot type resonators are used in the resonator system, and the anode and cathode diameters, as well as their ratio, are considerably larger than those used in previous designs for the same operating voltage and magnetic field. These features of the design

resulted from a systematic study of the causes of and possible cures for the
maximum current limitation encountered in long wavelength magnetrons.
The difficulty had been particularly noticeable in the 4J21-30, which on
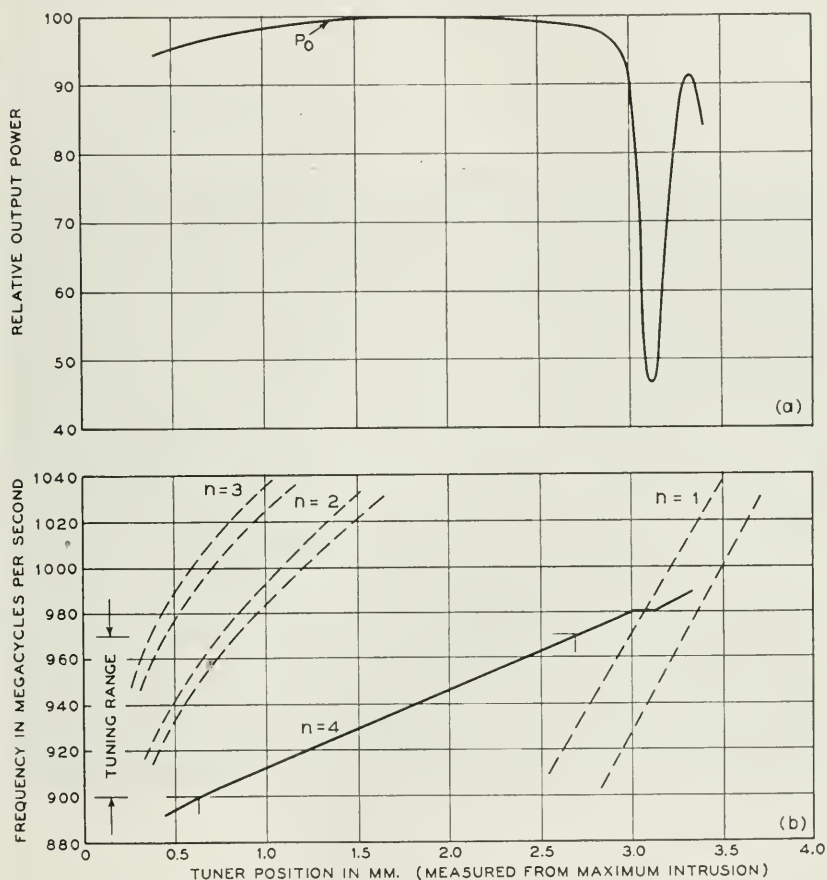
Fig. 56—Typical experimental curves of power output and mode frequency variation
with tuner position in the 4J51 magnetron. The following points of interest are to be
noted: the relatively more rapid tuning for modes of small n than for the π mode ($n = 4$);
the drop in output power at the coincidence of the frequencies of the π mode and one of the
components of the $n = 1$ mode; the limitation on the tuning range in the π mode imposed
by the crossing of mode frequency curves of other modes; the fact that the $n = 1, 2$, and
3 modes are doublets and the π mode a singlet.

hard tube modulators were able to pass the required current before π mode
failure with little margin. For some reason the performance of the early
experimental models of the tunable replacement was considerably inferior
in this respect and would not meet the specifications for which the magnetron
was being designed. Thus a study of the phenomenon was imperative. It

resulted in the 5J26 magnetron. It dictated the nature of the interaction space and resonator system and thus the cathode size and output coupling loop. Initial experiments were performed with a vane type resonator system [see Fig. 21(b)] having channel straps and an interaction space of approximately the size used in the 4J21-30 series.

From studies of dynamic V-I plots on the experimental tunable models and other magnetrons under a variety of operating conditions, it became clear that the phenomenon had to do with the rate of buildup of RF oscilla-
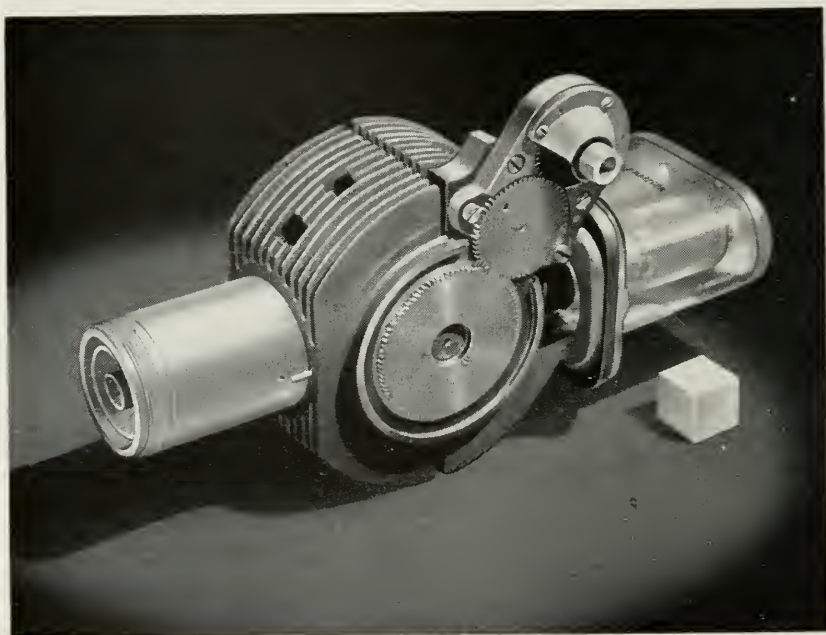
Fig. 57—The 5J26 tunable magnetron (600 kw., 1220 to 1350 mc/s)—the tunable replacement for the 4J21-30 magnetrons.

tion in the π mode and with the rate of rise of the applied DC voltage. The "mode skip" shown in the sequence of V-I plots of Fig. 38 and discussed in PART I is an example of the behavior of one of the experimental models used in this work when failure to oscillate in the π mode occurs.

Thus at constant magnetic field there appears to be a maximum DC voltage above which π mode oscillation is impossible. This voltage is presumably that beyond which the mean angular velocity of the electrons around the interaction space becomes enough greater than that of the traveling field component for the phase focusing to fail to maintain synchronism. The term "maximum current limitation" is thus in a sense a misnomer as Fig. 38(b) indicates. Here, π mode oscillation fails even to start and does not

fail at a maximum current on each pulse. The point at which π mode oscillation fails thus has to do with the relation of the rate of rise of DC voltage through the range of permissible values to the rate of buildup of RF oscillation. As one attempts to drive the magnetron harder by applying greater peak voltages to it, the rate of voltage rise in this region becomes greater, finally exceeding that which the rate of RF voltage buildup will permit.
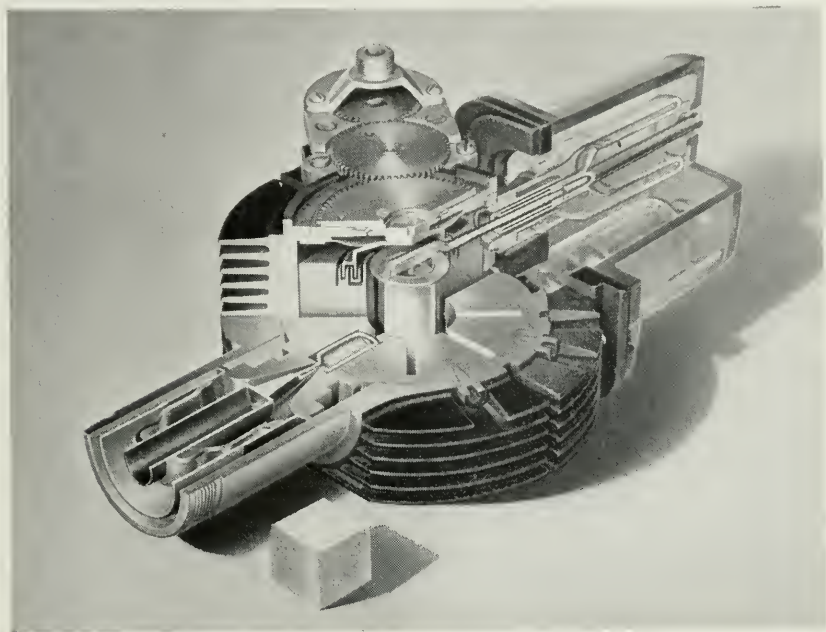
Fig. 58—A view of a sectioned 5J26 tunable magnetron (600 kw., 1220 to 1350 mc/s). Note that the tuning member, consisting of two concentric rings, is moved up and down inside the two recessed channel straps by a mechanism like that in the 4J51 magnetron (see Fig. 55). Compare the interaction space geometry and shape of resonators with those of the fixed frequency 4J21-30 magnetrons (Fig. 51), the changes having been made to achieve better starting characteristics as explained in the text. Note also the enlargement of the output resonator to accommodate the coupling loop, the RF chokes incorporated in the cathode support leads, and the large input lead construction to prevent external flashover.

The phenomenon of π mode failure was studied as a function of five parameters:

1. magnetic field, B;
2. frequency, f;
3. ratio of cathode to anode radii, $\dfrac{r_c}{r_a}$;
4. load conductance, G_s ;
5. rate of DC voltage rise, $\dfrac{dV}{dt}$.

The results were found to fit well into the picture of starting time behavior discussed in PART I. Thus, increase of frequency and decrease of load both result in a greater rate of RF voltage buildup and permit a greater rate of rise of DC voltage through the range where oscillation starts. The magnetron may thus be driven harder and more current passed. In agreement with these results, changes in the rate of voltage rise applied by the pulser, accomplished by changing its characteristics, also permitted oscillation to continue to higher current and voltage values. The dependence on B and r_c/r_a presumably are to be accounted for by the dependence of $G_e(V_{RF})$ on these quantities. It appears that $G_e(V_{RF})$ increases with both B and r_c/r_a.

Empirical relations of the critical current, I_c, and the electronic efficiency, η_e, at which π mode failure occurs, to the load conductance, G_s, were obtained.[25] Whereas I_c decreases with increasing G_s, η_e decreases.

Since the work was done at different frequencies and interaction space geometries, it was found convenient and instructive to express voltages, currents, and conductances as reduced variables of the type introduced by Slater, that is, as ratios to the voltage, V_o, and current, I_o, at the intersection of the Hartree line and the cutoff parabola, and the conductance, $G_o = \dfrac{I_o}{V_o}$, respectively. I_o and V_o may be determined from the expressions for the cut-off parabola and the Hartree line [expressions (8) and (16) of PART I, respectively]. Plotted in terms of the reduced variables I_c/I_o and G_s/G_o, it was found that the data relating I_c/I_o to G_s/G_o and η_e to G_s/G_o predicted approximately the same relationship independent of frequency and r_c/r_a. Thus, quite independent of the fundamental significance of these particular reduced variables to the functional dependence of the quantities involved, it appeared that they would be useful in the design of a new magnetron whose design parameters would lie in or near the range for which data were available.

The purpose of the work was to produce a tunable magnetron which would operate satisfactorily to currents in excess of those needed to meet the power output specifications. It was hoped to do this by redesign of the magnetron itself. This course rejected as unsatisfactory the alternative of limiting the rate of voltage rise on each equipment with which the magnetron would be used. The general goal of the proposed changes in magnetron design were to increase the rate of buildup of RF oscillation while at the same time keeping the electronic efficiency at a reasonably high value. These goals are not independent but oppose one another as has been seen. Thus, if r_c/r_a is increased, I_c increases, but η_e decreases. If G_s is increased to increase η_e, I_c drops. The hope of success in the venture lay in the prob-

---

[25] G_s was determined from the relation $G_s = \dfrac{Y_{OC}}{Q_L} = \dfrac{1}{\omega_L Q_L}$ [see equation (36) of PART I], using a calculated value for the total resonator inductance and the measured value of the loaded Q.

ability that these effects do not exactly cancel one another and that a net improvement in performance could be effected.

The design of the 5J26 was accomplished in the following way: The slot type resonator was chosen as it possesses a low value of inductance needed to give a high value of G_s for a given value of loaded Q.[25] The number of resonators was chosen to be eight, the anode length of the 4J21-30 retained, and the design carried out at $Q_L = 150$ and $f = 1220$ mc/s. The empirical relations used were that relating I_c/I_o to G_s/G_o, that relating η_e to G_s/G_o, and the equation of a constant B line on the performance chart obtained for the 4J21-30. In addition to the definitions of G_s, V_o, I_o, and G_o, there were to be met the operating conditions of $V = 27$ kv., $I = 46$ amps., and $B = 1400$ gauss, as well as a maximum value for the over-all diameter of the resonator system dictated by interchangeability. The value of resonator inductance was calculated for a terminated parallel plate line. With the choice of $I_c = 60$ amps., the above mentioned conditions could be solved numerically to give a definite set of design parameters. The anode and cathode radii, resonator length, and thus unstrapped wavelength were determined. The straps were designed to achieve the proper frequency and to provide sufficient tuning range with the tuning means employed. The ratio r_c/r_a was determined as 0.546, considerably in excess of the value 0.375 in the 4J21-30 magnetrons.

The output circuit of the 5J26 magnetron was similar to that used in the other tunable magnetrons of long wavelength. It employed the broad band design of external coupling and transition section to the coupling loop. The size of the loop was dictated by the coupling required to provide a Q_L of approximately 150. It may be seen in Fig. 58.

The input leads of the 5J26 magnetron were of new design, incorporating chokes in the leads to prevent leakage of RF energy picked up on the cathode, and large size center conductor to minimize RF voltage breakdown on the input leads.

The significant design and operating data for the 5J26 magnetron are included in TABLE I. This magnetron design provided a single tunable replacement of superior performance for the entire 4J21-30 series. In Fig. 59 are shown curves of mode frequency tuning throughout the π mode tuning range, indicating that in the strap tuning method difficulties with mode frequency interference are avoided.

15. MAGNETRONS FOR WAVELENGTHS NEAR 10 CENTIMETERS—UNDER 200 KILOWATTS

15.1 *The 706A-C and 714A Magnetrons:* The work in the wavelength region near 10 cm., described in Section 12, was continued in the study of variations from the British design. For the most part these early studies

had to do with the electron interaction space and the resonator system. Later, when driving equipment became available, life studies were started. In the main this work showed the British design to be a very nearly optimum as an unstrapped magnetron. Consequently the series of magnetrons
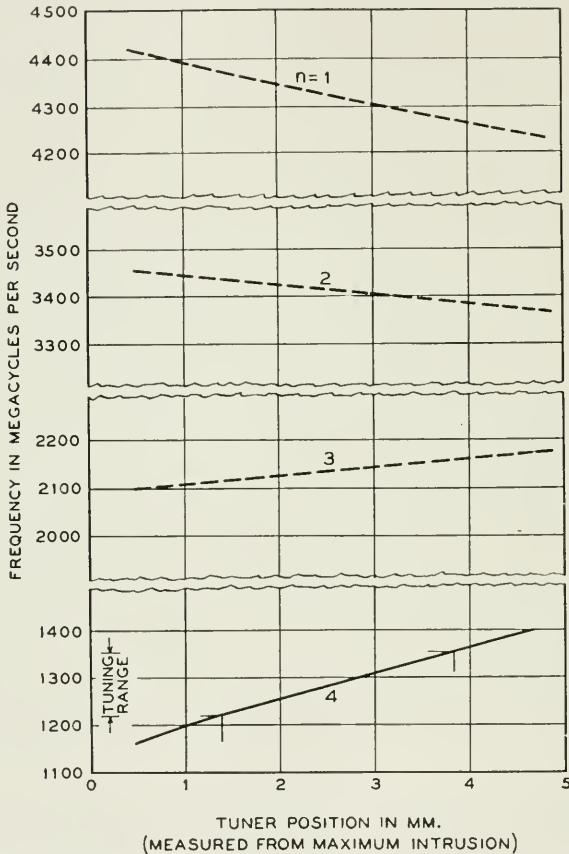
Fig. 59—Curves of mode frequency variation with tuner position in the 5J26 magnetron. The curve for $n = 4$ is experimental, those for $n = 1, 2,$ and 3 have been calculated from an equivalent circuit theory of the resonator system (doublet structure not shown). Note the large mode frequency separation and the complete freedom of the π mode from interference by other modes (compare Fig. 56).

having wavelengths near 10 cm. which were coded for American manufacture were similar to the British prototype. The experiments and calculations which had been made constituted a body of valuable information to be used later in departing from the original design.

The variations in the resonator system of the British magnetron were made primarily for the purpose of studying the parameters affecting frequency of operation. Magnetrons were built incorporating such variations as dimensional changes in the hole and slot resonators; the use of other forms of resonators such as slot resonators; and variations of factors affecting the coupling between adjacent resonators, as, for example, the volume of the end space region between the ends of the resonator system and the end covers of the magnetron. A series of magnetrons was built having from four to ten resonators, including odd numbers, but the majority of the experimentation was done with eight resonators as in the original system.

Besides affecting frequency, most of these resonator changes markedly affected the electronic operation as well. Undoubtedly this was the result of changes occurring in the mode frequency distribution. At the time, the significance of this distribution to magnetron operation was not fully appreciated nor, indeed, was it known for these magnetrons. As discussed in PART I in connection with curve (a) of Fig. 25, it was later found that the frequencies of several modes including the π mode were very nearly equal, making the electronic operation quite sensitive to the exact nature of the frequency distribution. This made the experiments difficult to interpret. None of the resonator systems, involving only changes in anode length and diameter with corresponding changes in the cathode, appeared to be appreciably better than that in the British magnetron.

In one rather important experiment the cathode diameter, and thus r_c/r_a, was varied to determine the value for optimum efficiency using the British anode dimensions. The result, confirmed later by comparison with British work, showed the original dimensions to be very nearly optimum.

Some variations of the output coupling to the resonator system were also tried. In one such variation the coupling loop was placed in the end space of the magnetron between two resonators. In this position the loop was coupled magnetically by the flux linking the two adjacent resonators, as well as directly by virtue of the fact that the end of the loop is fastened to the anode segment (see the discussion of output coupling in PART I). It will be recognized that this type of coupling is that used later in the 700A-D magnetrons. The output circuits in the first American 10 cm. magnetrons were like those in the British design.

The constructional techniques used in the British magnetron were followed with some variations. The most bothersome technique was that of making the vacuum seals between the end covers and the body of the magnetron. This was done by means of tin plated gold ring seals between the members. Little difficulty was encountered with this seal in production, however, and it was used throughout the production at the Western

Electric Co. of 10 cm. magnetrons under 200 kw. output power.

The first unstrapped magnetrons coded for manufacture by the Western Electric Co. were the 706A-C for frequencies near 3060 mc/s (9.8 cm.). The frequency differences were achieved by small changes in the resonator slot widths. Another magnetron, the 714A, oscillating at 3300 mc/s (9.1 cm.) was also coded at this time. It differed from the others in having slightly smaller resonator holes. These magnetrons were used in some of the earliest American ship and airborne radar systems. A set of typical operating data is included in TABLE II.

The improvement in measured over-all efficiency to approximately 25 per cent was undoubtedly the result of improvements in the technique of operation and of measuring the output power. Unlike the measurement of frequency, for which good techniques were already available, the measurement of output power initially was crude and unreliable. At first, estimates of power were based on the heating of resistors to incandescence, with no assurance that all the power was being dissipated in this load. The production of corona has always been spectacular evidence of RF power. Later, when water load techniques were used, all the power could be absorbed and measured. Means of line tuning permitted the adjustment of the magnetron load impedance to the point of maximum output power even though the impedance itself was unknown.

Some preliminary studies of magnetron tuning were made as auxiliary experiments. With one or more extra coupling loops terminated externally in adjustable coaxial tuners, tuning ranges of 2 to 4 per cent were achieved.

15.2 *The 706AY-GY, 714AY, and 718AY-EY Magnetrons:* A significant improvement in the multiresonator magnetron was the method of strapping, used by the British as a means of achieving mode frequency separation (see discussion in PART I). When used in the resonator system of the 706A-C magnetrons, the early British "mode locking straps" increased the π mode wavelength from 9.8 cm. to 10.7 cm. and resulted in an increase in operating efficiency by about a factor of two with much more freedom from "moding" difficulties. Here was a really worthwhile advance and no time was lost in making use of it. The resultant strapped magnetron, oscillating at 10.7 cm. wavelength, was used directly as the basis for a new family of magnetrons of wavelength near this value. These were coded as the 718AY-EY series. The cathode, output circuit, and general mechanical features of the 706A-C were adopted essentially without change. Experiments like those done on the original British design, when repeated on strapped magnetrons, indicated the design parameters still to be close to optimum. The cathodes used were plain, oxide coated, nickel cylinders.

Table II

Magnetrons for Wavelengths Near 10 Centimeters

	706A-C Unpackaged	714A Unpackaged	706AY-GY Unpackaged	714AY Unpackaged	718AY-EY Unpackaged	720A-E Unpackaged	4J45-47 Unpackaged
N	8	8	8	8	8	8	8
r_c (in.)	0.119	0.119	0.110	0.100	0.119	0.107	0.107
r_a (in.)	0.315	0.315	0.295	0.270	0.315	0.283	0.283
h (in.)	0.788	0.788	0.788	0.788	0.788	1.576	1.576
Magnet gap (in.)	1.480	1.480	1.490	1.490	1.490	2.750	2.750
Weight (lb.)	2.1	2.1	2.1	2.1	2.1	4.9	4.9
Resonators	hole and slot	hole and slot	hole and slot	hole and slot	hole and slot	hole and slot	hole and slot
Unstrapped λ (cm.)	9.8	9.1	~8.9	~8.3	9.8	~8.0	~8.0
Straps	none	none	British wire	British wire	British wire	double ring	double ring
λ (cm.)	9.8	9.1	9.8	9.1	10.7	10.7	10.7
f (mc/s)	3100–3019	3320–3280	3100–2914	3320–3280	2890–2720	2890–2720	2855–2750
Nearest mode	$n=3$	$n=3$	$n=3$	$n=3$	$n=3$	$n=3$	$n=3$
λ separation (%)	−1	−1	−11	−11	−11	−10	−10
Q_0	>1500	>1500	~1000	~1000	~1000	1500	1500
Q_{ext}	<100	<100	~120	~120	~120	130	130
η_c (%)	~93	~93	~89	~89	~89	93	93
Output circuit	coaxial	coaxial	coaxial	coaxial	coaxial	coaxial	coaxial
V (kv.)	11.0	11.0	11.4 18.5 20.9	11.6 17.7 19.5	11.9 17.9 20.0	21.0 24.0 27.0	21.0 27.0 23.0
I (amps.)	12.5	12.5	12.5 16 20	12.5 16 20	12.5 16 20	45 53 65	45 65 45
B (gauss)	1300	1300	1300 1900 2100	1500 2100 2250	1300 1800 1950	2300 2600 2900	2300 2900 2550
τ (μs)	1	1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 5
pps	1000	1000	2000 1000 1000	2000 1000 1000	2000 1000 1000	1000 650 500	1000 500 120
P_0 (kw.)	25	22	40 146 206	40 125 165	45 136 193	550 800 1100	550 1100 650
η (%)	18	16	28 49 49	28 47 42	30 47 48	58 63 63	58 63 63
η_n (%)	~19	~17	~31 ~55 ~55	~31 ~53 ~47	~34 ~53 ~54	62 63 63	62 65 63
PF (mc/s)	~15	~16	11 11 11	12 12 12	11 11 11	10 10 10	10 10 10

No particular precautions were taken to protect the active coating. Even so, lives both in the laboratory and in the field were shown to be well over 2000 hours. Much later, the mechanical structure of the cathode and its support was improved and a heavy glass protective housing designed to cover the input leads. The performance chart of Fig. 17 in PART I is that of one of the 718AY-EY series. Other data concerning these magnetrons are given in TABLE II.

Quite early in the study of the multicavity magnetron it had been realized that the mode frequency distribution of the resonator system is very important to efficient operation. Evidence had accumulated which indicated that something needed to be done to suppress all but one mode of oscillation. Attempts at influencing mode frequency separation were made in which fluxbarriers or partitions, either complete or partial, were used in the end spaces of the resonator system. These included what was essentially the distortion of the structure into the form of the so-called "serpentine" mentioned in PART I. Although striking changes in operation were produced, supporting the basic contention that there are more ways of doing a thing wrong than right, worthwhile improvements were not achieved.

Following the confirmation, here and elsewhere, of the British results with strapping and the coding of the series of 718AY-EY magnetrons, intensive work was started in the general investigation of the modes of the magnetron resonator system and their relation to the various possible methods for strapping. This work involved the determination of the mode patterns for mode identification, the technique for which was consequently developed. Various strapping schemes were devised, built into magnetron resonator systems, and studied. Among these were the echelon strapping also used by the British, single and double ring strapped systems, and diametral straps which were adaptable to external tuning of the magnetron. The importance of strap shielding and strap assymmetries discussed earlier were brought out in these studies. The effects of end space volume and anode length on mode frequency distribution for given strapping systems were also studied.

Strapping made possible greater tuning ranges with tunable magnetrons. Efficient magnetrons having broken echelon strapping on one end and diametral straps on the other, connected through a glass seal to an external coaxial tuner were built. Tuning ranges of over ±5 per cent were achieved in this way. There was no demand at the time for tunable magnetrons.

Following the introduction of the 718AY-EY magnetrons, their properties were reproduced at an extended range near the wavelengths of the 706A-C and at the wavelength of the 714A. The designs for these were arrived at by scaling the 718AY-EY to the appropriate wavelengths. The principles

of scaling discussed in PART I had been studied in connection with the extension of magnetron frequencies into the 20 to 45 cm. and 3 cm. wavelength ranges and were fairly well understood by this time. These new strapped magnetrons were coded as the 706AY-GY and the 714AY. The frequency variation over the 706AY-GY band was accomplished by variation of the resonator slot width. The characteristics of these magnetrons, apart from frequency, are essentially the same as those of the 718AY-EY series (see TABLE II). External and internal views of the 706AY-GY type are shown in Figs. 60 and 61. The strapping scheme to be seen in Fig. 61 is the early British system depicted schematically in Fig. 24(a).
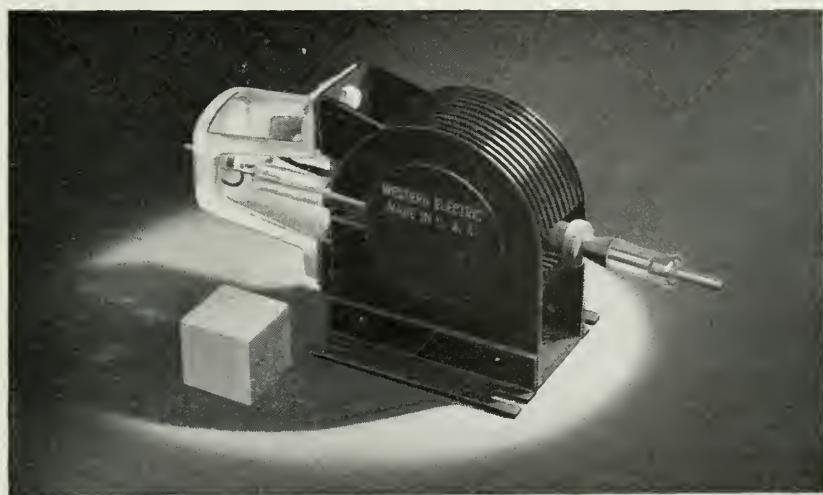
Fig. 60—An external view of a 706AY-GY magnetron (150 kw., ∼3000 mc/s).

The means of fabricating the anode block in manufacture is also of interest. This had to be inexpensive and fast. As worked out by Western Electric engineers, it consisted of boring the interaction space hole first and turning the end spaces into the solid copper block. This was followed by two operations in a multiple spindle drill press in which the eight resonator holes are drilled to size. Following this, a broach consisting of eight tapered series of projecting cutting edges, spaced equally around the main shaft, was drawn through the interaction space hole, cutting the eight slots to size in a single operation.

Although more powerful magnetrons were subsequently developed, these 10 cm. magnetrons of power output below 200 kw. are still considered to be highly satisfactory designs. Together with similar designs in Great Britain, they have been produced by several manufacturers in large quantity and

have seen extensive use in many forms of pulsed radar systems.   They have also served as the basis for much of the subsequent magnetron design work.

16. Magnetrons for Wavelength near 10 Centimeters— 200 Kilowatts to 1 Megawatt

16.1   *The 720A-E Magnetrons:*   A natural outgrowth of the early magnetron experimentation was the effort to generate higher power.   Since the earliest magnetrons were unstrapped there was much to be hoped for in
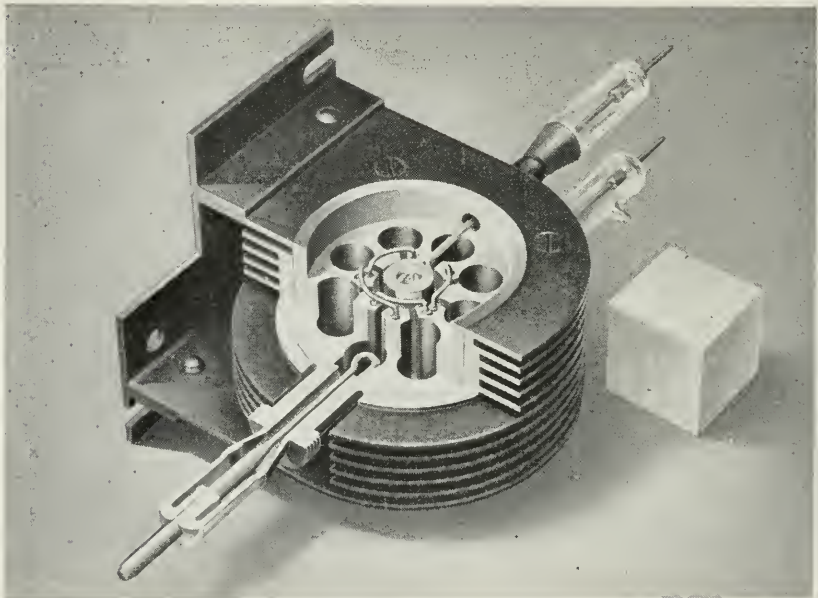
Fig .61—An internal view of a 706AY-GY magnetron (150 kw., ∼3000 mc/s).   Note the type of wire strapping [compare Fig. 24(a)] and the simple coaxial output circuit.

improved efficiency.   Higher power magnetrons would have to be designed to operate at voltages above 20 kv. and currents of greater than 50 amps. To accomplish these objectives, the interaction space was enlarged, in some designs involving a greater number of resonators, and the anode length increased by two or even three times.   Although it was possible to develop over 200 kw. with magnetrons of eight, ten, and twelve resonators, efficiencies were poor, seldom exceeding 20 per cent at maximum loading.   This was even poorer than unstrapped 10 cm. magnetrons like the 706A-C.   The reduction in mode frequency separation attendant upon increases in anode length and number of resonators was later found to be the cause.

Following their first use in the 718AY-EY magnetrons, straps were in-

corporated in the high power designs with immediate improvement in performance. The output circuits used were simple loop to coaxial lead combinations like those in use on the 706A-C.

Radar needs for higher power near 10 cm. had developed to the point of a specific requirement for a magnetron of not less than 350 kw. output. Since it was believed that the best one could obtain from the 718AY-EY was 250 kw. even at the risk of shortened cathode life, the requirement made necessary a new design. In the work on higher power magnetrons, better than 300 kw. had been obtained at efficiencies greater than 50 per cent at maximum loading with what was essentially a double length 718AY-EY
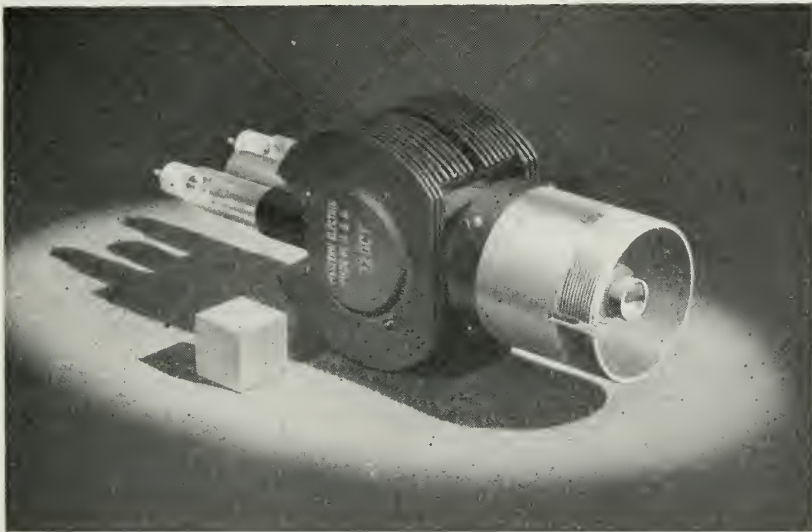
Fig. 62—An external view of a 720A-E magnetron (1000 kw., ~2800 mc/s).

type. Thus it appeared that the new power requirement could be met with such a design. The press of time made it necessary to concentrate further efforts on improving this design, with not much if any time to be devoted to further experimentation along other lines. The chief drawback of the double length magnetron was the large magnet it required. Several months later an equally efficient magnetron, shorter, and having a larger number of circuit elements might have been produced, but the original program had advanced beyond the stage at which it could be changed. The improvements in design of the double length 718AY-EY type magnetron resulted in the 720A-E series.

An external view of the 720A-E magnetron is shown in Fig. 62. A cutaway model is shown in Fig. 63 in which may be seen the nature of the

resonator system, straps, cathode, and output circuit.   In Table II are given dimensions and other data concerning these magnetrons.

The resonator system has eight hole and slot type cavities.   Its anode length is 4 cm., double that of the 706AY-GY.   It is strapped with double ring shielded straps of special design which could be stamped out of sheet copper and oven brazed in position in a wide ledge trepanned around the interaction space into the ends of the anode block.   These straps were easier to control in dimensions than the echelon straps, used in early experimental models, which had made necessary the development of a pretuning
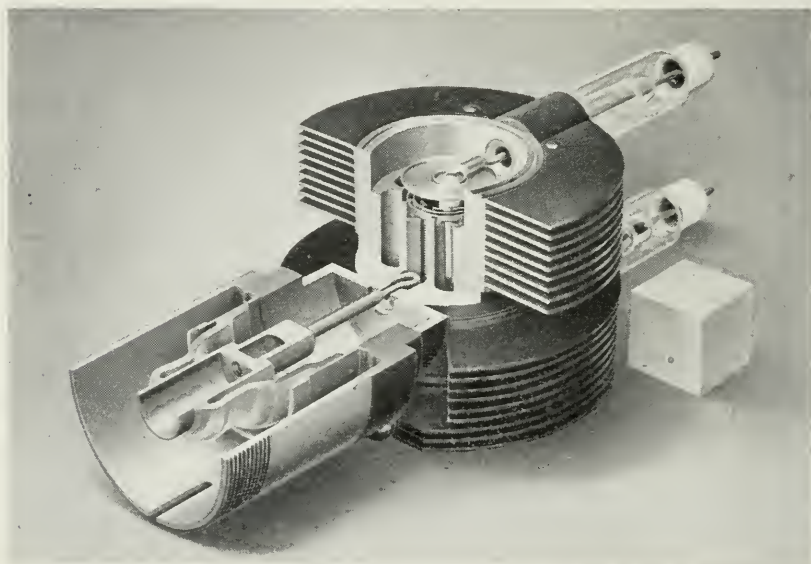
Fig. 63—A view of a cut-away 720A-E magnetron (1000 kw., ∼2800 mc/s) showing the large cathode end disks and the impedance transformer sections in the output circuit.

technique.   The large number of wire straps used had made it impossible to predict from visual inspection that a given anode would eventually oscillate at a specified wavelength.   The straps were adjusted manually to bring the frequency of resonance determined before sealing and pumping into the required range.   These heavily strapped models operated at efficiencies up to 65 per cent, but their initial operating wavelengths were near 13 cm. The high operating efficiency was achieved at the required wavelength by scaling the structure.   The increase in magnetic field required by the scaling was achieved without increase in magnet size by incorporation of steel disks in the end covers of the magnetron.   The new stamped straps required considerably less manipulation during pretuning than the echelon straps.

The five wavelength groups of the 720A-E employ straps of different height.

The cathode was similar to that of the 706AY-GY but longer and with different end disk design. The surface was a plain oxide coating on the nickel cylinder base. Adequate life under the most stringent of its operating conditions was obtained (see Table II). The radical departure in cathode end disk design was necessitated by the pickup of RF energy by the cathode. This was aggravated by the length of the cathode. An end disk of conventional size, being closer to the inner strap than to the outer strap, is more influenced by its potential. Since the inner straps are π radians out of phase and the cathode approximately a half wavelength long, conditions were quite right for considerable induction of RF potential on the cathode. Considerable RF power was radiated by the cathode leads. Measurements on a non-oscillating magnetron indicated that cathode pickup could be neutralized very nearly by increasing the end disk diameter. If the end disk completely covered both straps, the charge induced on it by one strap was essentially balanced by that induced by the other strap. The 720A-E with large cathode end disks radiates very little RF energy from its cathode leads.

The greater power which the 720A-E produced made it necessary to redesign the output circuit to include the transformer inside the vacuum. Voltages in the external transformer had been sufficiently high to break down the line. The redesign was carried out empirically by a substitution method. A replica of the output circuit was constructed and terminated in the impedance required to load the magnetron properly. In this replica the loop was replaced by a coaxial line and standing wave detector through which CW power was fed. In this way one could determine the impedance at the loop terminals necessary for proper loading. Then with a specially constructed output circuit, in which the dimensions between the glass seal and the loop terminals could be varied, the dimensional changes necessary to achieve the required impedance at the loop with match in the output load coaxial was determined. The dimensions determined in this way provided the final design. The 720A-E was one of the early "preplumbed" magnetrons.

A coaxial lead coupling using chokes instead of contacts was developed for the output circuit and used not only at 10 cm. but on all the later magnetrons of longer wavelength as has been described. The centrifugal glass seal at the outer conductor is to be noted (see Fig. 63). The original tungsten center conductor to which an external "thimble" was soldered (similar to that in the 5J23, see Fig. 49) was replaced by a Kovar cup to which the glass was sealed directly. Greater strength, better cooling, and elimination

of the high voltage gradient in air which had existed at the small diameter tungsten rod resulted.

In Fig. 64 is shown a typical performance chart for the 720A-E series. Other performance data are given in TABLE II. In Fig. 44 is shown one of the 720A-E magnetrons mounted in its magnet.
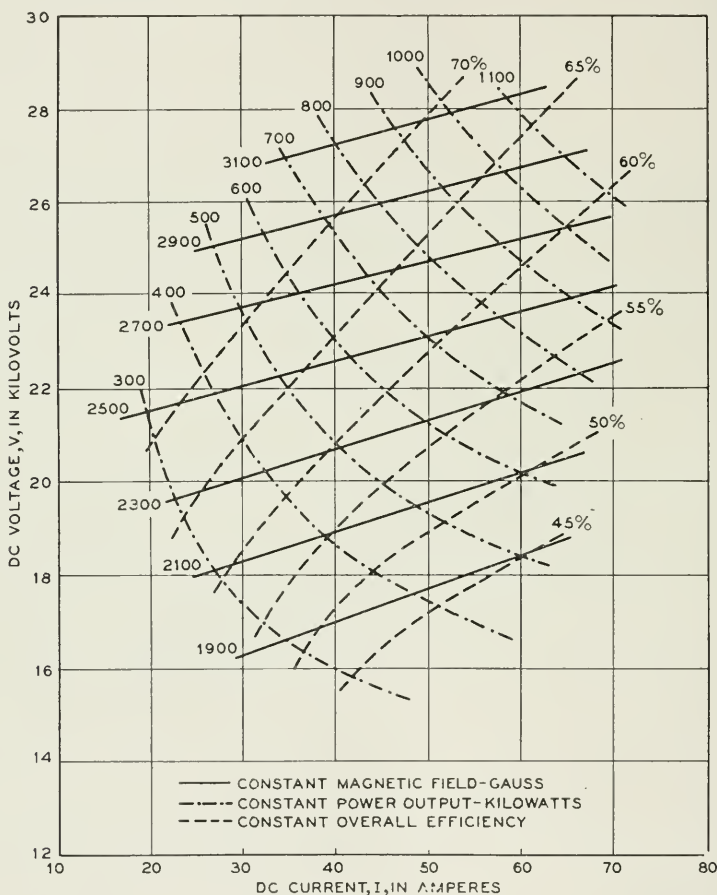
Fig. 64—A typical performance chart of the 720A-E magnetrons (~2800 mc/s).

In the use of the 720A-E on line type modulators the difficulty with what was then called voltage "overshoot" was first encountered. It is the result of the finite buildup time for the RF oscillation during which the pulse voltage rises to values higher than required for π mode operation. It is particularly aggravated on line type modulators for which the magnetron, before it oscillates and draws current, presents a mismatch to the line network causing the voltage to rise to a much higher value. The difficulty was

circumvented by connecting across the modulator network a condenser in series with a resistance equal to the characteristic impedance of the network. The condenser is of low impedance for rapid voltage changes, terminating the network properly during the voltage rise. The condenser, however, stores some energy which must be dissipated at the end of the pulse. The discharge of this energy through the magnetron causes very weak oscillations to occur following the main pulse. This condition somewhat lengthens the interval of time following the pulse during which the radar receiver may not detect an echo.

16.2 *The 4J45-47 Magnetrons:* For the extension of radar range it had been found desirable to use pulses of longer duration with consequent increase in total energy radiated per pulse. As indicated in PART I, the longer the pulse, the narrower the frequency spectrum. A long pulse consequently makes possible a narrower receiver band width, an improved receiver signal to noise ratio, and greater range. Although it was necessary to reduce the peak power somewhat in operating the 720A-E at a pulse length of 5 microseconds, a substantial increase in range could still be effected. The Radiation Laboratory at M. I. T. proposed to use 720A-E type magnetrons on long pulse operation and performed many tests of these magnetrons under these conditions. Many 720A-E magnetrons performed satisfactorily under these long pulse conditions. However, arcing both internally between cathode and anode and externally across the input leads deteriorated the performance and made the life expectancy of the magnetron questionable.

To meet the new demands a new series of magnetrons, the 4J45-47 was developed by making two design changes in the 720A-E. The plain, oxide coated cathode was replaced with a cathode of the "mesh" type to be described in Section 21. MAGNETRON CATHODES. The "mesh" cathode proved to be more rugged, involved less arcing, was longer lived, and required a minimum of changes in internal structure.

External arcing at the input leads was minimized by lengthening the glass section, accomplished by increasing the over-all lead length and shortening the metal section of the lead at the magnetron body, and by enclosing the leads in housings filled with a silicone anticorona compound. These changes were limited by the requirement of interchangeability with the 720A-E.

The modified 720B-D magnetrons were coded the 4J45-47. Operating characteristics are given in TABLE II. These magnetrons were used with long pulses and in the usual service for which the 720A-E had been designed.

17. MAGNETRONS FOR WAVELENGTHS NEAR 3 CENTIMETERS— UNDER 100 KILOWATTS

17.1 *The 725A and 730A Magnetrons:* As the antenna size of a radar is reduced, it is necessary to reduce the operating wavelength to maintain

the same width of radiated beam and resolving power. To meet the needs for aircraft and submarine radar and other applications in which antenna size must be small, shorter wavelength magnetrons were essential. The development of magnetrons of frequency near 10,000 mc/s (3 cm. wavelength) was part of the expanding program of magnetron research and development which grew out of the earliest experiments. The work was concentrated at 3.2 cm., roughly a factor of three below the initial 10 cm. work. From the earliest 3 cm. magnetron developments was evolved an efficient 3.2 cm. magnetron, the 725A, requiring roughly the same driving conditions as were used in the first 10 cm. magnetron applications.

An operating 3 cm. magnetron was built in the summer of 1941. The crude techniques then in use in the 10 cm. wavelength range were adapted to the 3 cm. range, and 5 kw. peak power from this magnetron was measured in a coaxial water load. This design involved an unstrapped resonator system having eighteen quarter wavelength slots. The choice of so many resonators was made in order that large cathode and anode dimensions could be used, later shown to be unreasonable for the voltage range employed. In addition the design suffered from a confusion of many modes of the resonator system and from an inadequate output circuit.

A later design, similar in some respects to one upon which work was done at the M. I. T. Radiation Laboratory, made use of a resonator system having twelve slots. This design operated in a mode other than the π mode, probably that for which $n = 3$ or $n = 4$. The M. I. T. version, the 2J21, was the first 3 cm. magnetron to be manufactured in quantity. Its power output was about 15 kw., generated at an efficiency of from 12 to 15 per cent. Many variations of this design were made without major success in an attempt to obtain efficient and reliable operation.

Prior to the introduction of straps, attempts were made to adapt 10 cm. magnetron designs to 3 cm. While none came up to the 10 cm. operating efficiencies, one such, having an unstrapped resonator system of eight hole and slot resonators, was used extensively in systems experimental work in our Laboratories.

By this time it was realized that some fundamental reason existed for the failure of these 3 cm. magnetrons to reach efficiencies comparable with those obtained at 10 cm. The resonator systems were unstrapped at 3 cm., and they operated in a mode other than the π mode. However, this did not fully explain their failure. Attempts were made to operate strapped 3 cm. magnetrons in the π mode, but they still did not result in the desired or expected improvements. The trouble lay not only in the mode frequency distribution of the resonator system but also in the size of the interaction space. Prior to this time, 3 cm. magnetrons were made with larger inter-

action spaces than direct scaling from 10 cm. would indicate. This had been done to achieve greater cathode size and smaller operating magnetic field. It appeared that 3 cm. magnetrons had been suitable only for operation at high voltage and high magnetic field and were being operated far below their range of maximum efficiency or even of reasonable operation. A decisive step forward was taken when a strapped magnetron was built which was rigorously scaled, according to the scaling principles discussed in PART I, from 10 cm. designs of the same operating voltage and current as that desired at 3 cm.

The first such magnetrons were scaled from the 10.7 cm. strapped magnetrons having eight resonators, the 718AY-EY, with only minor changes being made in the resonators for mechanical reasons. The results obtained were very encouraging although they were erratic for a variety of reasons. Mechanical construction was difficult and not reproducible before newer tools and techniques for making and assembling small parts were introduced. Output circuit variations in many cases completely masked good electronic operation, and it was not until a carefully considered and executed study of the output circuit design was instituted that consistent results were obtained. However, the results were such that it was decided to shift all emphasis in 3 cm. magnetron development at the Bell Laboratories to strapped designs scaled from 10 cm. magnetrons.

This was by no means an easy decision to make. For example, it meant the use of very small parts and clearances, of order 0.010 in., and very close tolerances. That such a magnetron would be feasible for large scale production was by no means obvious. A cathode, 0.100 in. or less in diameter, which must deliver a considerably higher current density than any previous magnetron cathode, would be necessary. Whether such a cathode, at the expected operating conditions, would have any appreciable life was not known. Furthermore, a scaled 3 cm. magnetron would require a much higher magnetic field than previous 3 cm. designs, a demand which might increase the magnet weight.

On the other hand, it was possible that the improved electronic efficiency of the scaled magnetron would result in less rigorous treatment of the cathode in spite of its small size. In spite of the increase in magnetic field, the decrease in anode length of the scaled magnetron made possible the reduction of the magnet gap to a point where it was conceivable that no greater magnetomotive force and size of magnet would be required than in previous 3 cm. designs. Greater stability and freedom from mode troubles could be expected.

Early strapped magnetrons having eight resonators had anode diameters of order 0.175 in. and cathode diameters of order 0.065 in. Magnetrons of

over-all efficiency better than 45 per cent at a pulling figure of 15 mc/s were made.

The cathode life in these magnetrons was very short—10 to 30 hours. Reduction of the severity of the demands made of the cathode was necessary. Strapped magnetrons having twelve resonators and larger cathodes were designed and a few built with promising results.

The 3 cm. magnetron development program was later broadened by the inclusion of personnel from the NDRC Radiation Laboratories at the Massachusetts Institute of Technology and at Columbia University. Effort on magnetrons having twelve resonators was expanded and prosecuted vigorously along with the work on the eight resonator types. Somewhat later the effort was concentrated solely on the twelve resonator magnetron. The design which was made shortly thereafter became, with only minor changes, the 725A.

Experience gained with previous magnetrons at all wavelengths enabled one to design a mechanically feasible resonator system with proper strapping to give the desired RF characteristics and π mode resonant frequency. The interaction space, on the basis of previous work and Hartree's oscillation condition, was designed to meet the specification of 12 kv. and 12 amps. input. The design was complicated both mechanically and electrically by the requirement of interchangeability with the 2J21. Thus good operation at 10 kv. and 10 amps. input was necessary.

The resonator system of the 725A is machined into a separable "anode insert" which, when completed, is brazed into position in the so-called "shell" carrying the leads and cooling fins. The straps are placed in a channel for shielding from the interaction space and are broken on one end.

The first cathodes were the ordinary, nickel sleeve type with double carbonate coating sprayed onto its surface. Not long after the start of the 3 cm. magnetron development, it was recognized that the cathode development problem would be a large part of the over-all project. Indeed, the 725A proved to be an excellent magnetron for cathode development work. Many of the cathode improvements and new designs developed for it have been used successfully in other magnetrons. The cathode as finally developed for the 725A was a nickel blank, complete with end disks and heater chamber turned out of nickel rod. Over the cylindrical portion was welded or sintered a fine nickel mesh in the interstices of which the active coating was applied. This is shown in Fig. 79 and will be described more fully with other cathode developments in Section 21. MAGNETRON CATHODES. Under normal operating conditions the cathode heater is turned off, the necessary heat being provided by electron back bombardment.

Considerable effort was expended in the design of a satisfactory and repro-

ducible output circuit for the 725A. The requirement of interchangeability with the 2J21 restricted the output wave guide flange in its position relative to the axis of the anode and the mounting flange. Previous 3 cm. magnetrons having hole and slot resonators utilized coaxial output circuits coupled to the resonator system by a loop located in the median plane of one of the resonators as in Fig. 1. Although it was recognized that a wave guide output for a small magnetron of this type would be more elegant and reproducible than a coaxial output, difficulty in twisting the plane of polarization from the resonator to the output wave guide flange, involved in the interchangeability requirement, and the lack of experience with such outputs made necessary the choice of an output for the final design similar to the original. To obviate the removal of a large portion of one resonator to accommodate the output and to make the loop easily accessible for inspection and adjustment, it was placed in the end space directly over the circular hole of one of the resonators. This so-called "halo" type loop is shown in Fig. 66. The coaxial line was terminated, as in previous 3 cm. magnetrons, in a junction to wave guide, fabricated as part of the magnetron. The output circuit was designed around a matched junction which, it was hoped, would have the best electrical breakdown strength at the reduced pressures of high altitude operation, would be frequency insensitive, and would avoid adjustment of the junction on individual magnetrons. The entire antenna of the junction was enclosed in the vacuum envelope of the magnetron as may be seen in Fig. 66.

The method of attack on the output circuit problem from this point on may be summarized as follows: First, a matched junction from coaxial to wave guide such that it could be applied to the final over-all design, had to be achieved. Then, for a given choice of loop size and position relative to the resonator, it was necessary to obtain the transformer properties of the coaxial line immediately adjacent to the loop so that the proper over-all coupling from magnetron resonator to wave guide load was achieved. Here again the mechanical restriction on the distance between the resonator and wave guide axes was found to be cumbersome. Two loop sizes were tried, that requiring the higher standing wave in the coaxial transformer to couple it properly being later discarded. At any given stage in the development, the procedure consisted of finding the impedance to be presented to the loop by the coaxial output. This was done by transforming through the entire output circuit to the loop terminals, the impedance needed in the output wave guide for proper loading, using the transformer parameters measured on a simulated output circuit. This impedance at the loop terminals was obtained by variation of the transformer properties of the coaxial line between the loop terminals and the coaxial to wave guide junction. The

change of diameter of the coaxial center conductor at a critical distance from the loop terminals is to be seen in Fig. 66.   The height of the output loop above the resonator end surface was a convenient variable upon which the over-all output coupling depended.   This was not without disadvantages, however, since its value for reproducibility had to be held to within 0.001 in. Its nominal height above the anode is 0.011 in.

While the scaling of 10 cm. magnetrons to 3 cm. demanded higher magnetic fields than had previously been used at 3 cm., the magnet size was
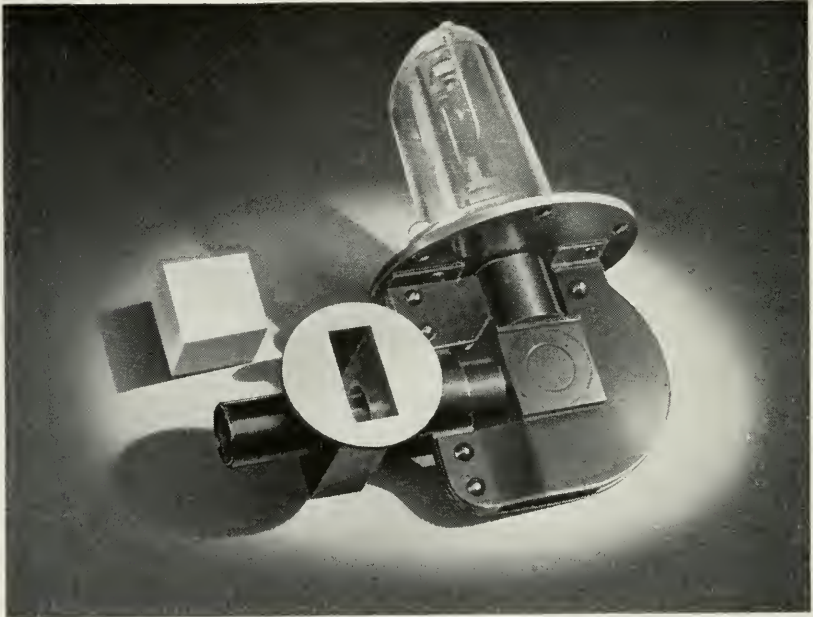
Fig. 65—An external view of the 725A magnetron (55 kw., 9375 mc/s), showing the attached coaxial to wave guide junction in the output circuit.

kept down by the reduction of the magnet gap made possible in the scaling of anode length.   A further reduction in effective magnet gap was made by the inclusion of steel disks in the end covers of the magnetron.   With these changes, the magnetomotive force required to supply the field for the 725A was almost identical to that required by the 2J21 and no extensive magnet redesign was necessary.

The requirement of interchangeability with the 2J21 made it necessary to place the input and output leads of the 725A at right angles (see Fig. 65). Along with the 725A, there was developed a version, coded the 730A, with input leads mounted directly opposite to the output in a so-called "straight

TABLE III

MAGNETRONS FOR WAVELENGTHS NEAR 3 CENTIMETERS

	725A, 730A Unpackaged	2J48-50 Unpackaged	2J55-56 Packaged	2J51 Packaged Tunable	4J50, 4J78 Packaged	4J52 Packaged
N	12	12	12	12	16	16
r_c (in.)	0.051	0.051	0.062	0.062	0.104	0.104
r_a (in.)	0.102	0.102	0.125	0.125	0.159	0.159
h (in.)	0.250	0.250	0.250	0.250	0.250	0.250
Magnet gap (in.)	0.625	0.625	0.384	0.384	0.380	0.380
Weight (lb.)	1.5,1.1	1.5	3.8	4.6	9.5	5.5
Resonators	hole and slot	hole and slot	hole and slot	hole and slot	hole and slot	hole and slot
Unstrapped λ (cm.)	2.3	~2.3	~2.3	~2.6	2.46	2.46
Straps	double ring	double ring	double ring	double ring	double ring	double ring
λ (cm.)	3.2	3.22, 3.3, 3.4	3.2, 3.24	3.33	3.2, 3.3	3.2
f (mc/s)	9375 ± 30	9315 ± 5, 9080 ± 80, 8825 ± 75	9375 ± 30, 9245 ± 55	8500 to 9600	9375 ± 30, 9080 ± 80	9375 ± 30
Nearest mode	$n = 5$	$n = 5$	$n = 5$	$n = 5$	$n = 7$	$n = 7$
λ separation (%)	−16	−16	−16	~−16	−14	−14
Tuning	—	—	—	resonator inductance	—	—
$\Delta\lambda$ (%)	—	—	—	12.1	—	—
Tuner travel (in.)	—	—	—	0.110	—	—
Q_0	680	680	680	630-460	850	850
Q_{ext}	280	280	290	290	350	350
η_c (%)	71	71	70	65	71	71
Output circuit	coaxial-wave guide	coaxial-wave guide	coaxial-wave guide	coaxial-wave guide	wave guide	wave guide
V (kv.)	10.0 12.0 13.0	10.0 12.0 13.0	12.0	11.0 12.0 14.3	22.0 22.0	15.0 15.0
I (amps.)	10 10 12	10 10 12	12	11 12 14	27 27	15 13.5
B (gauss)	4500 5400 5650	4500 5400 5650	3350	3050 3350 4000	6900 6900	4950 4950
τ (μs)	1 1 1	1 1 1	1	1 1 1	1 5	1 5.5
pps	1000 1000 1000	1000 1000 1000	1000	1000 1000 1000	500 200	1000 200
P_0 (kw.)	31 44 56	31 44 56	50	33 40 60	280 280	110 95
η (%)	31 37 36	31 37 36	35	27 28 30	47 47	49 47
η_c (%)	44 52 51	44 52 51	50	42 43 46	66 66	69 66
PF (mc/s)	13.5 13.5 13.5	13.5 13.5 13.5	12	12 12 12	12 12	12 12

through" design.   Experimental models were released for systems tests and the need developed for such magnetrons to be used in some special radar systems.   The 730A is identical to the 725A except for external mechanical details.   An external view is shown in Fig. 67.

Details of the final design of the 725A (and 730A except for the mechanical differences noted) may be seen in Fig. 66.   Data on these magnetrons and their performance are given in TABLE III.   A typical performance chart
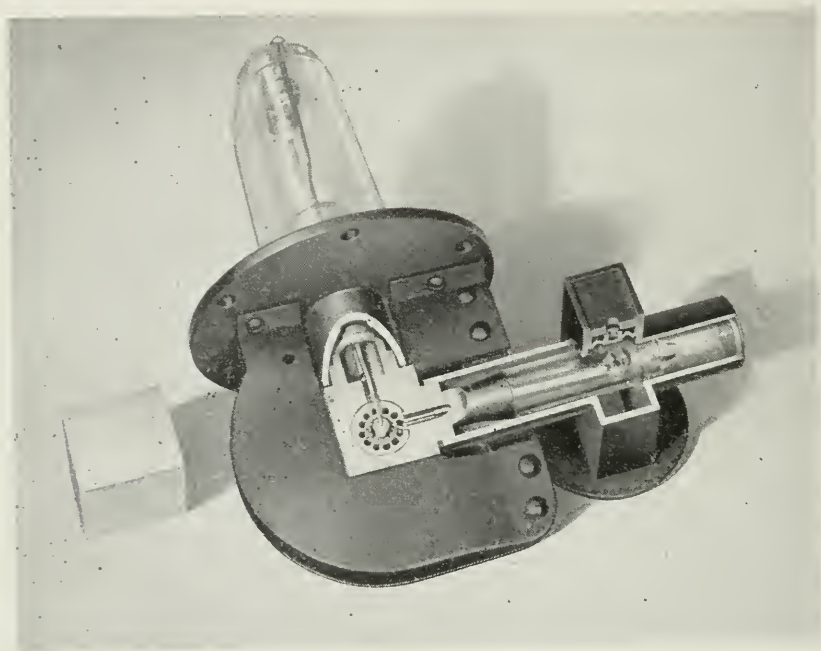
Fig. 66—A view of a cut-away 725A magnetron (55 kw., 9375 mc/s).   Note the breaks in the double ring straps, the so-called "halo" loop above the hole of the output resonator, the step in the center conductor of the output coaxial providing the necessary impedance transformation, and the details of the coaxial to wave guide junction.

is shown in Fig. 68.   Experience has indicated that it is entirely adequate for 150 kw. input at duty cycles up to 0.001 and, with later cathode designs, at pulse lengths up to 5μs.   The RF spectrum is satisfactory even under rather extreme conditions of mechanical vibration and shock, and the operation under a variety of conditions is remarkably free from "moding".   Difficulties have been encountered in attempts to push its power input with 5μs pulses much above 175 kw. although with short pulses at lenient duty cycle operation with as much as 300 kw. peak input power has been achieved.

Many difficulties were encountered in the early stages of production above those of training personnel to handle small parts and to perform new types of operations.   Maintenance of cathode centering through all operations had to be rigidly supervised since off-center cathodes resulted in poor operation.   The size and position of the output loop were very sensitive variables directly affecting output power and pulling figure.   The latter were also marked functions of variables in the transformer of the output circuit.   It was found, for example, that variations in the glass, and particularly in the bead supporting the center conductor, were causing considerable
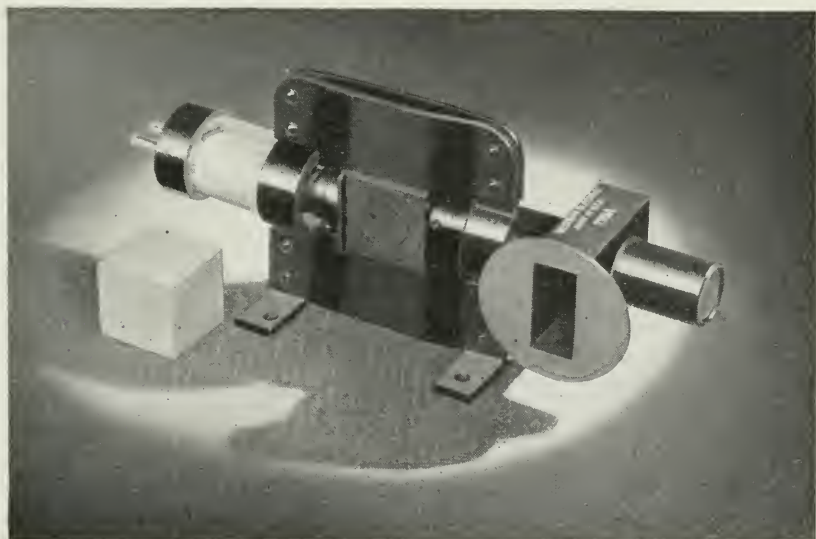
Fig. 67—The 730A magnetron (55 kw., 9375 mc/s)—the "straight through" version of the 725A.

spread in characteristics.   Closer tolerance on the glass, involving the use of a molded bead, greatly improved this situation.

With the very narrow limits specified for the operating frequency of the magnetron, it was essential that the anode pretuning be done very precisely. Here, considerable difficulty was encountered with a pretuning arrangement in which the indication of resonance depended on the equivalent line length of the output circuit between the anode block and the detector in the wave guide.   Variability in this electrical length caused the actual resonance to vary over a considerably wider range than the required 9375 ± 30 mc/s. By use of a detector connected at the output loop it was possible to reduce the percentage of magnetrons eliminated for being outside the frequency

limits at final test from more than 50 per cent to less than 1 per cent. At one point in the development, when it appeared that the frequency spread could not easily be reduced, a tuning screw mechanism was designed with which the frequency of the resonator system could be varied by movement
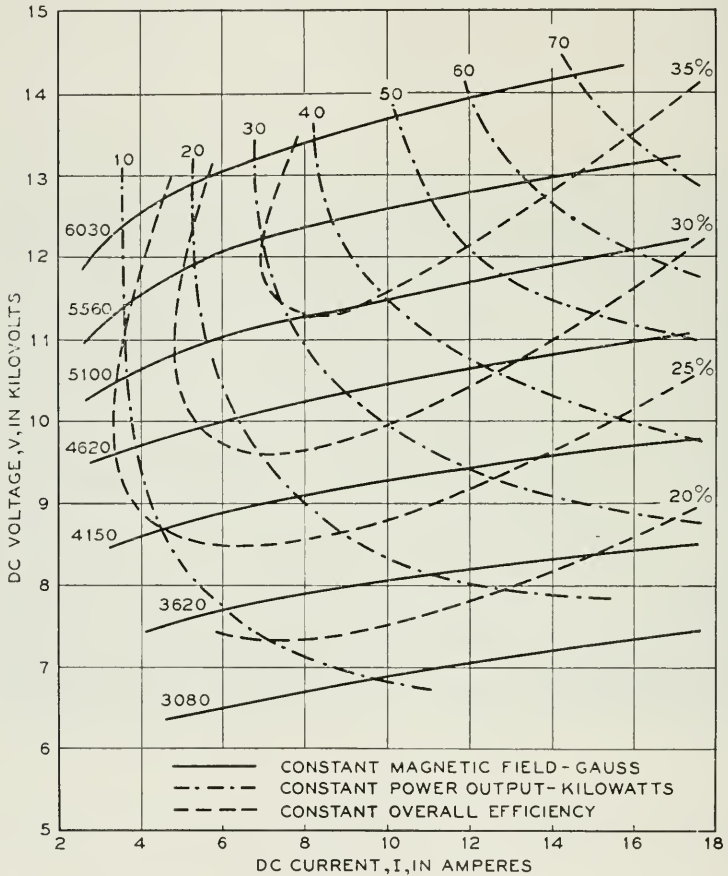
Fig. 68—A typical performance chart of the 725A and 730A magnetrons (9375 mc/s).

of a plug through the side wall of one resonator and which could be adjusted from outside the vacuum envelope at the final acceptance test. Although the mechanism was built, tested, and found to operate satisfactorily, the improvement of the pretuning technique made its adoption unnecessary.

During the initial stages of production it was verified that the pulling figure of the 725A and 730A could be adjusted within limits by variation of the position of the end plate shorting the wave guide section into which the

coaxial output circuit of these magnetrons is terminated.   It is possible by a displacement of 0.040 in. of the whole end plate to effect about a 6 mc/s change in pulling figure near the value of 15 mc/s.   An end plate consisting of a convoluted diaphragm was constructed which could be adjusted at final test if necessary.   By this means it was possible to make the adjustment without disturbing the airtight seal needed in some pressurized installations. The adjustable end plate is to be seen in the cutaway model of Fig. 66.

Despite initial difficulties, it was possible in the first nine months of production to increase the yield of shippable magnetrons meeting all the test specifications from a very low initial figure to over 85 per cent of all those completed and reaching final acceptance tests.   Although subsequent tightening of the test specifications, such as the increase of pulse length at test from 1μs at 1000 pps to 2 μs at 325 pps, caused temporary setbacks in the percentage yield, it was nevertheless increased until it ran consistently better than 95 per cent.

Although better magnetrons have since been built in the 3 cm. wavelength range, the 725A occupies a special place in magnetron development at these wavelengths.   It represents the first high efficiency, π mode, strapped magnetron for 3 cm. operation, and as such has served as the prototype of most of the subsequent 3 cm. magnetron developments in the United States and Great Britain.

The 725A was manufactured by the Western Electric Co. in this country and the Northern Electric Co. of Canada.   The Raytheon Mfg. Co. used Bell Laboratories' design information to produce a 725A magnetron but redesigned the resonator system to use a vane type more adapted to their manufacturing practice.   The total number of 725A magnetrons produced in these three plants was over 300,000, indicating the extent to which it was used during the war.

17.2   *The 2J48, 2J49, 2J50, and 2J53/725A Magnetrons:* Following completion of the 725A magnetron at 3.2 cm., requests were made for three similar magnetrons.   These were to have the same characteristics as the 725A, differing only in their frequencies of operation.

The 2J48 is an exact duplicate of the 725A but has a narrower spread in frequency, $\pm$ 5 mc/s, around a mean value of 60 mc/s lower than the nominal 725A frequency of 9375 mc/s.   The 2J48 was never manufactured by the Western Electric Co.

The 2J49 and 2J50 are 725A types operating at 3.3 cm. and 3.4 cm., respectively.   These new wavelengths were met by a new resonator design involving larger resonator holes than in the 725A.   The nominal wavelength of the new design was 3.35 cm., tuned either into the 3.3 cm. or the 3.4 cm. band by strap manipulation.

In an effort to get high power from the 725A magnetron operating on long pulses, considerable work was done to build a cathode for a power input of 225 kw. While a few thousand such magnetrons were made and tested at the higher test condition, entirely satisfactory operation was never achieved. Magnetrons tested at 225 kw. input were coded the 2J53/725A but after the first production run were not manufactured again. The 4J50 and 4J52 magnetrons are now available for this range of operating conditions.
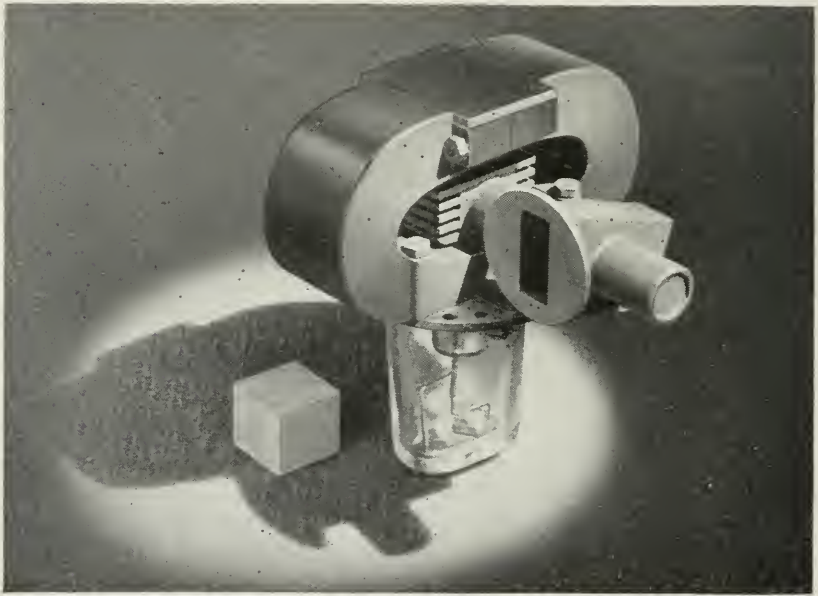
Fig. 69—The 2J55 magnetron (55 kw., 9375 mc/s)—the "packaged" version of the 725A. Compare with the size of the 725A and its magnet shown in Fig. 44.

17.3 *The 2J55 and 2J56 Magnetrons:* The 2J55 and 2J56 are by-products of the development of a tunable 3 cm. magnetron, the 2J51, described below. The 2J51 design, being "packaged" and thus lighter and more compact than the 725A and its magnet, with which it is interchange-able, was ideally suited for conversion into a fixed frequency, "packaged", 3 cm. magnetron. The only design changes required were those associated with the removal of the tuning mechanism and the change in dimensions of the resonator system needed to bring the wavelength to the proper value. The two magnetrons designed in this way for operation at the wavelengths 3.2 and 3.25 cm. were coded the 2J55 and 2J56, respectively. Each weighs 3 lbs. 12 oz. as compared to the weight of 11 lbs. 2 oz. of the 725A mag-netron and its magnet.

A photograph of the 2J55 magnetron is shown in Fig. 69 and operating data are included in TABLE III.

18. TUNABLE MAGNETRONS FOR WAVELENGTHS NEAR 3 CENTIMETERS

18.1 *The 2J51 Magnetron:* As in the 20 to 45 cm. wavelength region, the demand arose for tunable magnetron replacements when fixed frequency magnetrons had become available in the 3 cm. band. A significant step toward satisfying this demand was taken at the NDRC Radiation Labora-
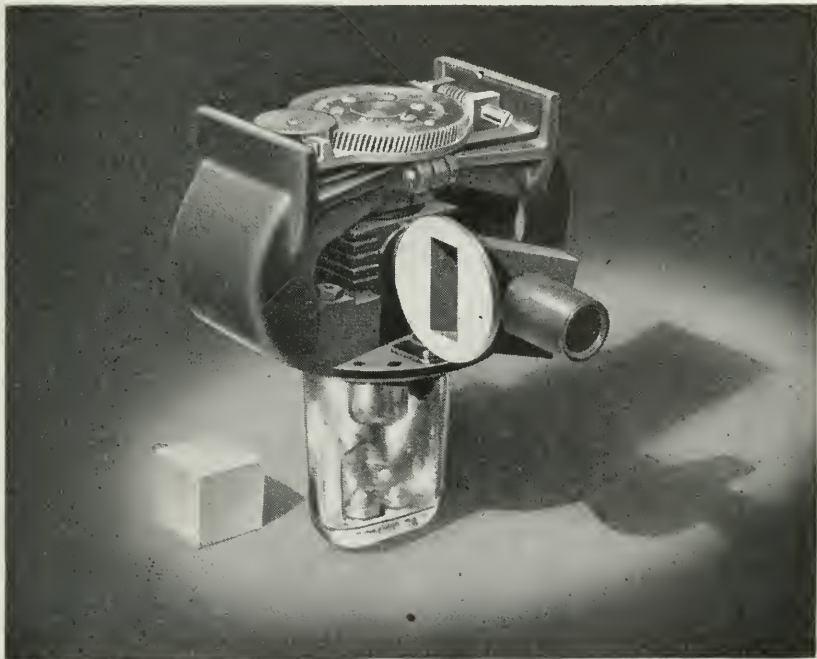
Fig. 70—An external view of the 2J51, tunable, "packaged" magnetron (55 kw., 8500 to 9600 mc/s).

tory at Columbia University in 1943 when a tuned, "packaged" magnetron was built around a 725A anode structure and output circuit. The tuning scheme used was that of variation of resonator inductance, as discussed in PART I, by a so-called tuning head consisting of a set of pins which project into the holes of the resonator system. The mechanism for driving the tuning head is contained inside one of the magnet pole pieces. The other pole piece carries the axial cathode mount. Two magnets are clamped to opposite sides of the pole pieces in the usual "packaged" construction. These and other details are shown in the photographs in Fig. 70 and 71 of complete and cutaway models of the magnetron, coded the 2J51, as it was put into production at the Western Electric Co.

From the inception of the project until the bulk of the work of developing the tunable 3 cm. magnetron was transferred to the Bell Laboratories, a number of problems had been encountered and solved at the Columbia Laboratory. It had been shown that tuning a 725A type magnetron was
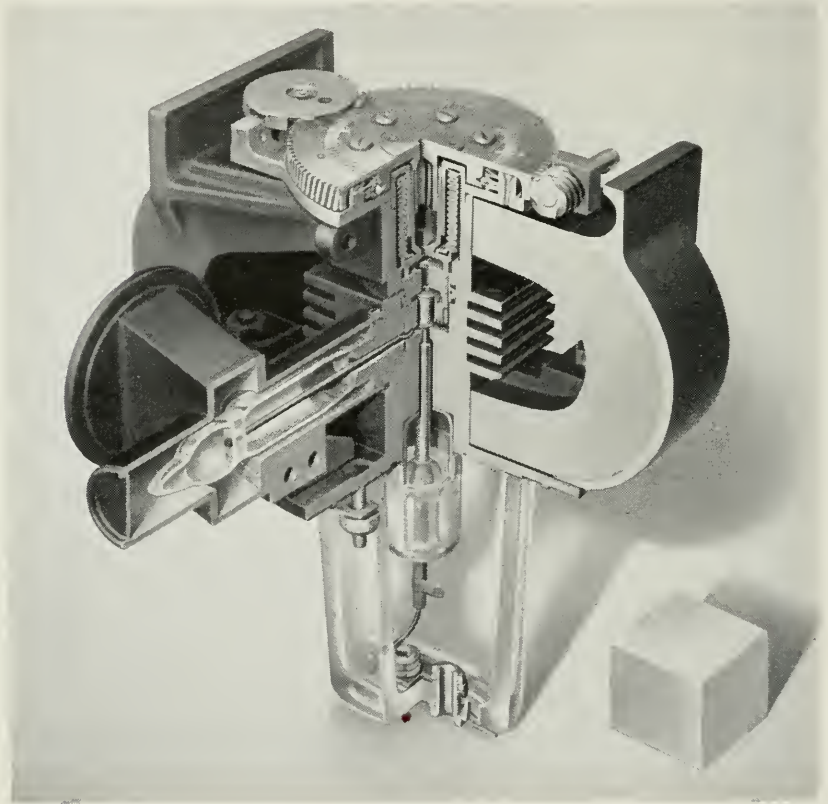
Fig. 71—A view of a sectioned 2J51, tunable, "packaged" magnetron (55 kw., 8500 to 9600 mc/s). Features of interest to be seen are: the tuning pins, the drive mechanism and vacuum bellows, the axial cathode mount, the glass bead support of the center conductor of the coaxial output line, and the sheathed permanent magnet construction.

feasible by the means chosen. The range of wavelength from 3.13 to 3.53 cm. could easily be spanned. The variation in frequency was found to be very nearly linear with position of the tuning pins. It was shown how a spurious resonance of the pin structure occurring within the frequency range could be displaced so as to cause no difficulty. Although no work of revision had been done on the 725A output circuit used in the tunable magnetron, it was demonstrated that tolerably "flat" characteristics of

output power and pulling figure against frequency could be obtained with it. The fundamental mechanical problems of "packaging" the magnetron and providing facilities for driving the tuning head were shown to be feasible and a smooth working and resettable driving mechanism designed and constructed.

The Columbia Radiation Laboratory built a series of these magnetrons, several of which were used in the development of tunable radar systems. The Bell Laboratories had cooperated from the start by supplying 725A magnetron parts, and by making anode inserts of special dimensions to raise the maximum attainable wavelength of the resonator system. It was decided that the Bell Laboratories should take over the main burden of further development of the magnetron for manufacture by the Western Electric Co. By this time its specifications had been quite definitely established. It was to be a magnetron with the general power output capabilities of the 725A, tunable over the frequency range from 8500 to 9600 mc/s (approximately 3.13 to 3.53 cm.). It was to be interchangeable with the 725A magnetron in the sense that it could be installed directly in any radar system using the 725A when the magnetron and its magnet were removed. Interchangeability, as usual, was one of the most annoying requirements. The magnetron was to be "packaged" and provided with magnetic shunts so that the magnetic fields necessary for operation at 10 kv. and 10 amps., 12 kv. and 12 amps., and 14 kv. and 14 amps. could be attained with a single magnet design. No specific requirements on variation of output characteristics were made at that time, but it was understood that only a magnetron whose pulling figure varied by an amount of the order of 3 mc/s or less over the entire frequency band would be acceptable. Similarly, it was understood that the tuning mechanism should provide smooth variation in frequency with a minimum of backlash and of such resettability that the use of a wavemeter would not be required in setting the radar system to a new frequency.

During the course of the development at the Bell Laboratories it became necessary to redesign the magnetron in a number of important respects. The requirement that the magnetron operate at 14 kv. and 14 amps. demanded a magnetic field for which the magnet weight would be prohibitive. Consequently, anode and cathode radii were scaled from those of the 725A magnetron by the factor 1.2. The larger cathode resulting from this redesign made easier the problem of attaining the maximum of 200 kw. peak input power required by the specifications. However, it caused a recurrence of the difficulty with a resonance of the tuning pin structure appearing in the 8500 to 9600 mc/s range. This had been found at the Columbia Laboratory to be a resonance of the capacitance between the pins and anode structure

with the inductance of the end space region through which the pins project from the magnet pole piece into the resonators. As before, the frequency of this resonance was removed by filling the end space with a copper ring or collar.

In the course of study of this resonance and of output circuit characteristics, an extensive program of testing of non-oscillating experimental models, whether operable or not, was carried out to supplement the data of oscillation tests. From these measurements, three important types of data were obtained, namely, the variation of pulling figure with frequency, the amount of RF energy loss introduced by the insertion of the tuning pins, and the variation of frequency with position of the tuning head. The first of these data, combined with measurements of the output circuit transformer characteristics made on simulated and adjustable models, yielded a complete understanding of the functioning of the output circuit. The effect of coupling loop size and of the variation of the dimensions of the coaxial to wave guide junction on the performance of the magnetron over the frequency range were determined. Although other designs of output circuits of the general 725A type were experimented with, it was possible to adjust the parameters of the original 725A output circuit to give a reasonably flat characteristic with frequency. In Fig. 72 are shown the variation of pulling figure and Q_{ext} over the frequency band.

The variation of unloaded Q with frequency, obtained in the above mentioned non-oscillating tests, indicated that stainless steel pins or even such pins copper plated to a considerable thickness were unsatisfactory. It was found that under the heat treatment occurring during brazing operations the copper plate and steel diffused through each other so as to increase markedly the surface resistance of the pins. Copper sheathed pins were found to be satisfactory. Solid copper was finally used as it was found to possess sufficient strength under mechanical shock test. Even so, as Fig. 72 indicates, the Q_0 of the resonator system falls off with increasing frequency and pin penetration.

Determination of the variation of frequency with position of the tuning pins dictated the design of the drive mechanism to provide the necessary range of motion. It was found that, with a pin diameter of 0.064 in. moving in resonator holes of 0.088 in. diameter, the frequency band could be spanned in a total travel of 0.110 in. Fig. 72 shows the nearly linear relationship between frequency and pin penetration.

The drive mechanism of the early Columbia models was redesigned to make possible its fabrication by a single brazing operation in a jig and to provide a sleeve rather than a thread bearing. The backlash achieved in the final design is such that one may reset the tuning head for a given frequency

to within 1 mc/s which corresponds to a displacement of the pins of 0.0001 in. and to 0.1 turn of the worm wheel drive. The tuning mechanism is provided with a positive stop at each end of the tuning range. The frequency band is covered in five revolutions of the main drive screw and gear. A flexible vacuum envelope is provided by means of sylphon bellows. Details of the design may be seen in the sectioned model shown in Fig. 71.
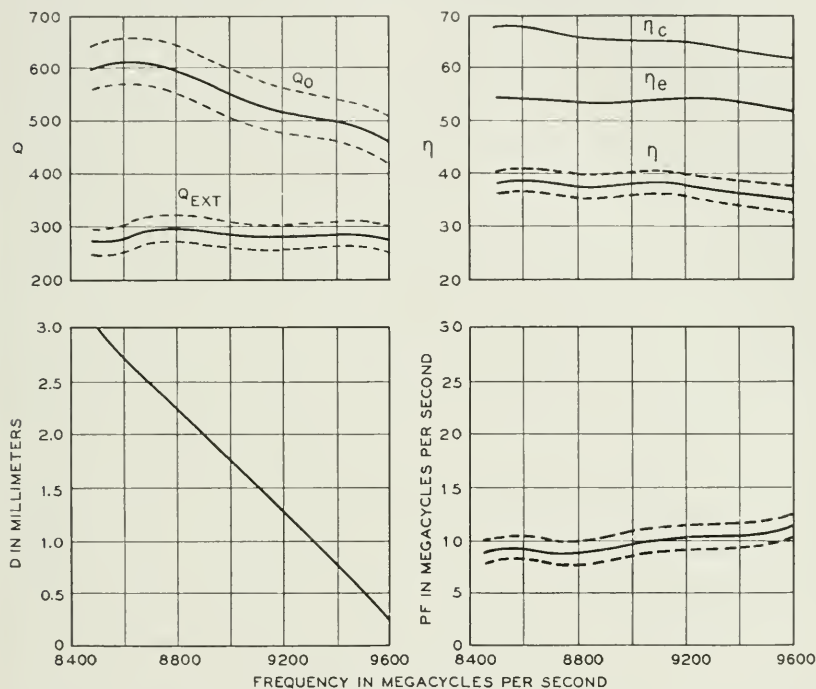
Fig. 72—Smoothed characteristics of the 2J51 tunable magnetron (55 kw., 8500 to 9600 mc/s) throughout its frequency band. Shown is the dependence upon frequency of the unloaded Q, Q_o; of the external Q, Q_{ext}; of the circuit efficiency, η_c; of the electronic efficiency, η_e; of the over-all efficiency, η; and of the pulling figure , PF. Also shown is the dependence of the operating frequency upon position of the tuning pins, D, measured from the position of maximum intrusion. The dashed lines on either side of the experimental curves indicate the range of possible variation of the data. The curves for η_c and η_e have been determined from the experimental results by the use of equation (25) of PART I and the relation $\eta = \eta_e \eta_c$.

The elimination of radial cathode support leads and the use of the magnetic pole pieces as the end covers of the magnetron body made possible a large reduction in the magnet gap and magnetomotive force required to produce the necessary magnetic field. Because of the requirement of interchangeability with the 725A magnetron, a very awkward V-shape magnet design in which the leakage flux was tremendous could not be avoided.

Fortunately, before production commenced, relaxation of the interchange-ability requirements made a change in the magnet form possible, and the much more efficient U-shape of those shown in Fig. 70 was adopted.

In performance, the 2J51 magnetron is essentially a 725A capable of operating over a 12 per cent frequency band. TABLE III gives two sets of operating conditions available through the use of the magnetic shunts. Fig. 72 shows several circuit characteristics of the magnetron over the frequency band. Also shown is the variation in over-all operating efficiency. The slight drop in efficiency observed as the frequency increases is due primarily to the decrease in Q_0. The electronic efficiency increases slightly with increasing frequency. The performance chart of the magnetron at its intermediate magnetic field is very similar to that of the 725A at the same field value, as shown in Fig. 68. The performance of the 2J51, when connected to a long, mismatched line exhibits a periodicity in power output, pulling figure, and frequency variation as the tuning is changed. This is expected in such a case because of the resulting changes occurring in the electrical length of the line with changing frequency. If the mismatch is sufficiently great, the condition may exist in which periodic regions of the frequency spectrum are completely unattainable, as discussed in section 10.2 *Frequency Sensitive Loads.*

The 2J51 represents an attempt to design considerable versatility into one magnetron. As has been described in the discussion of the 2J55 and 2J56, the development of the 2J51 provided the opportunity to make available "packaged" magnetrons of fixed wavelengths near 3 cm. by the omission of the tuning mechanism.

19. Magnetrons for Wavelengths near 3 Centimeters— 100 to 300 Kilowatts

19.1. *The 4J50, 4J52, and 4J78 Magnetrons:* In the later stages of development of the 725A, it had been demonstrated that a high efficiency scaled magnetron of wavelength near 3 cm. was feasible. It also had become evident that, freed from the hampering restrictions on input power and mechanical interchangeability, a magnetron of considerably greater output power capabilities could be made. The achievement of a magnetron design capable of delivering at least 200 kw. output power at a duty cycle of 0.001 or greater was set as a first objective. This was achieved with a good margin. During its entire course, the program was actively participated in by members of the Radiation Laboratory at M. I. T. working in residence at the Bell Laboratories.

The objective of 200 kw. peak output power at the factory test was not considered to be the maximum obtainable from a 3 cm. magnetron but

was largely determined by the state of development of systems and components when the work began. It was desirable to produce a magnetron which, while marking a very substantial gain over the 725A in power output, should at the same time be of reasonable weight, air-cooled, and capable of operation with a pulser of modest dimensions and of working into the system components in existence or under development.

The design began with the choice of a resonator system having sixteen resonators. The anode diameter was chosen for an operating voltage range of 20 to 25 kv. The increase in number of resonators from twelve to sixteen and operation at higher voltage necessitated a considerable increase in cathode diameter over that of the 725A. The length of the anode was left as before because it was felt that the desired increase in total cathode emission would be available from the increase in cathode area and that the magnet weight ought to be kept as low as possible.

Experimental models incorporating these design changes were built in the 725A type structure with radial cathode mount, "halo" coupling loop and coaxial output terminating in a junction to wave guide. The performance of these early models was satisfactory. Operating efficiency better than that of the 725A mgnetron was achieved over the range of currents from 4 to 40 amps.

However, it was easily to be seen that the 725A type structure was by no means the ideal for a 3 cm. magnetron of increased output power. The output circuit was marginal in its power handling capabilities; it could transmit no more than 300 kw. Furthermore, the opportunity presented itself of eliminating other troublesome features such as the "halo" loop and the coaxial to wave guide junction by designing a wave guide output whose critical dimensions are machined to size. The type of cathode structure of the 725A was limited in its heat dissipation in comparison with the axial type of cathode mount made possible by the use of magnet "packaging". In addition, the axial type cathode mount is superior from the standpoints of DC voltage breakdown and economy of space. Considerations of weight also favored the "packaged" structure, in which the magnet pole pieces form an integral part of the magnetron structure. Many of these first experimental models suffered from a drooping voltage-current characteristic at constant magnetic field in the region of low currents. This effect and its attendant loss of operating efficiency became progressively worse as the magnetron was operated. It was hoped that the parasitic electron emission from the cathode end disks, believed from experiments with non-emissive coating materials to be responsible for the effect, could be eliminated or reduced in an axially mounted cathode.

Accordingly, an entirely new design was undertaken, built around the

sixteen resonator anode structure used in the early work but including the features of "packaging," axial cathode mount, and wave guide output. A number of vexing but nevertheless interesting problems was encountered. Principal among these were the tendency to operate in another mode under certain conditions and the problem of obtaining satisfactory magnetic field uniformity in a design which necessitated the removal of a considerable portion of the center of the magnet pole piece to accommodate the axial cathode mount. Each of these difficulties was surmounted as is described in the detailed discussion later. The resonator system was redesigned quite late in the development to take advantage of the possibility of operating at a lower electronic conductance and hence higher RF voltage for the same output power. The advantages of this step were a reduction in the tendency of the magnetron to "mode" and an increase in unloaded Q.

The development program briefly sketched above resulted in three coded magnetrons, the 4J50, 4J52, and 4J78. The 4J50 and 4J78 magnetrons, identical except for frequency, were built with magnets large enough to permit operation near the maximum power capabilities of the design and with larger input leads to withstand the higher DC voltage required. The 4J52 magnetron, although incorporating the same internal design as the other models, was built with smaller magnets for operation at a set of conditions easily attainable with the higher power design but beyond the reach of the 725A. Developmental work on all three magnetrons was conducted simultaneously.

In the 4J50 magnetron a hole and slot resonator system was adopted. The remarkable freedom of the 725A magnetron from "moding" difficulties had made it appear desirable to retain as large a mode frequency separation as possible. Equivalent circuit theory indicated that to do this the increase in the number of resonators from twelve to sixteen would necessitate about twice the strap capacitance of the 725A. This would have required extremely wide straps for which the recess channel in the anode structure would be very difficult to trepan. For this reason the resonator system was strapped less heavily; the $n = 8$ and $n = 7$ mode frequencies differed by 19 per cent rather than 25 per cent as in the 725A.

In none of the early models built into 725A type structures was any "moding" or "misfiring" observed. When axial cathode and wave guide output were introduced, however, "moding" was experienced both in those with no strap breaks and those with the usual double break at one end of the anode. By enlarging the cathode diameter such that the ratio of cathode to anode radii increased from 0.60 to 0.66, the difficulty was removed in models with broken straps, but it was not possible to eliminate it in those with unbroken straps. Decreasing the r_c/r_a ratio increased the tendency to "mode."

It might be surmised that the trouble in the unbroken strap case was due to a component of $n = 7$ which did not couple to the output circuit. The phenomenon was studied in an operating magnetron with small electrostatic probes built into several resonators. The $n = 7$ components were identified, their relative intensities being approximately those expected. It is of some interest to note that the "moding" encountered here differed from that seen in magnetrons of longer wavelength. The magnetron seemed invariably to start in the π mode, falling into oscillation in the $n = 7$ mode more rapidly as loading increased. As might be expected from this behavior, the mode boundary on a performance chart is little affected by the rate of rise of the pulse as it had been in other magnetrons. Later evidence made it appear that the change in the operating voltage of the $n = 7$ mode brought about by the redesign of the resonator system was a decided improvement.

The necessity of using strap breaks led to an unforeseen difficulty. Under the influence of RF electric forces, the overhanging ends of the straps at the strap breaks moved together and shorted, causing failure after only a few hours of operation. The trouble was eliminated by removing the overhanging portions of the straps with no noticeably harmful effects.

In a 3 cm. magnetron, strapped as heavily as the 4J50, circuit losses become very important in determining the over-all efficiency. A noticeable improvement in unloaded Q was effected when an atmosphere of prepurified N_2 was substituted for the CO_2-alcohol mixture which had previously been used to prevent oxidation during the final brazing. The CO_2-alcohol method had been abandoned for another reason, namely, that chemical analysis showed carbon deposited on the steel pole-pieces underneath the copper plating.

If one determines the circuit efficiency for matched load by impedance measurements on a non-oscillating magnetron [see equation (25) of PART I], one may then calculate its value for any load. From measurements of over-all efficiency as a function of load conductance, the dependence of the electronic efficiency on this conductance may be determined. The curve of Fig. 19 was obtained in this way for the 4J50. The fact that the electronic efficiency, as seen in Fig. 19 is practically independent of load, that is to say, of electronic conductance, together with some observations of the load sensitivity of the "moding" in the 4J52, led to consideration of a new design for the resonator system. It was found invariably that the boundary for mode change in the 4J52 magnetron moved to higher currents as the load impedance was changed to values further removed from the frequency sink and power maximum in any direction on the Rieke diagram. This suggests that, as the circuit conductance [G_s of equation (36)] is

increased, the decrease in RF voltage in the π mode places that mode at a disadvantage in competition with the $n = 7$ mode.

As a result of these considerations, a new resonator system was designed in which the electronic conductance at the normal operating point was to be two thirds of that in the first design. To maintain the pulling figure invariant, it was necessary to reduce the total resonator capacitance. Although all of this capacitance could not be removed from the straps without reducing the mode frequency separation too drastically, a good proportion of it could be. From this, one might expect a gain in unloaded Q since the straps are the lossiest part of the circuit.

The new resonator system is more satisfactory than the old. Mechanically, it is neater in that the outer strap does not extend into the holes of the hole and slot resonators. A definite increase in over-all efficiency attributable to a greater Q_o was observed. The mode separation is 17 per cent, much greater than the expected value. An analysis of data on the $n = 8$, 7, and 6 mode wave lengths, by means of equivalent circuit theory, indicated that this is due to the straps being effectively shorter although their physical length is unchanged in the new design. This is plausible, however, for in the new design the outer strap is connected at a higher voltage point along the resonator.

The cathode structure of the 4J50, 4J52, and 4J78 magnetrons represents a radical departure from previous designs. It was desirable from the production standpoint to be able to build the cathode structure completely as a subassembly before mounting it in the magnetron. This entailed making holes in the magnet pole pieces large enough for the cathode to pass through; these holes were, in fact, made of the same diameter as the anode (0.319 in.). Since no radial cathode support leads were required, it was possible to reduce the end spaces to a height of 0.065 in., making the magnet pole gap 0.380 in. It was recognized that the large hole to gap ratio would result in two bad effects: first, a loss of magnetic field and, second, an antibarreling of the magnetic field which results in an axial force acting on the electrons, directed away from the center of the interaction space. Both of these difficulties were surmounted by the use of cathode end structures made of high Curie temperature permendur (50 per cent iron, 50 per cent cobalt).

The permendur pieces, toroidal in shape, are mounted at the two ends of the cathode (see Figs. 75 and 78). Their cross section and location, necessary for a nearly uniform field over 80 per cent of the gap and a focusing field over the remainder, were determined in electrolytic tank experiments. Since the permendur pieces fill up a part of the hole in the magnet pole piece and have a separation less than the pole gap, they contribute substantially

to the magnetic field in the anode-cathode region. The effective gap is reduced from 0.380 in. to 0.340 in. by their presence, resulting in about a 20 per cent decrease in magnet weight. While their primary function is magnetic, the permendur pieces also serve as normal cathode end disks preventing electrons from reaching the pole pieces. Their smooth contour gives a good DC voltage breakdown condition between anode and cathode. Finally, they are mounted in such a way that there is some thermal isolation from the cathode and that the migration path from cathode to external surface of the permendur is quite long. These features discourage electron emission from the cathode end structures.

Almost any degree of cooling may be obtained with an axially mounted cathode without the attendant disadvantages of the heavy tungsten leads which fill up the end spaces of a magnetron having radial cathode mounting. An upper limit to the cooling is set by the fact that the cathode must be raised to 1050°C in activation, using a heater which can be contained within the cathode sleeve without encroaching too much upon its wall thickness. The high temperature needed during activation sets certain limits upon the materials which may be used in the cathode structure. Since the cathode is mounted from one end only (this being dictated by assembly considerations), mechanical strength is exceedingly important. A cathode, once off center, is subjected to magnetic forces on the permendur ends which tend to pull it further off center. Fortunately, measurement at the magnetic fields used in normal operation shows that these forces are only of the order of one pound for a cathode position 0.015 in. off the axis, increasing to about 3.5 pounds when the permendur is in contact with the wall.

The first cathode surfaces used in these magnetrons were of the "mesh" type. Later, the newly developed, sintered nickel matrix type was used. More recently, considerations of strength have led to the introduction of molybdenum cathodes. The cathode assembly is brazed into a hollow metal cone in the base of which is a receptacle type heater and cathode connection. Because of the length of the supporting structure and the relatively high temperature at which it operates, there is considerable expansion and resultant motion of the cathode in the axial direction. The cathode is offset when cold to correct for this expansion. Performance is not very sensitive to cathode location, but in extreme cases of misalignment moding difficulties may be aggravated. The cathode structure is to be seen in Fig. 78 and is shown mounted in the magnetron in Fig. 75. Attention should be called to the external cathode input lead which is of heavy glass and Kovar construction sufficiently rugged to make unnecessary any protective housing.

In the wave guide output designed for the 4J50 magnetron, the necessary

impedance transformation is accomplished by a quarter wavelength section opening directly into the output wave guide at one end and into the outside wall of one resonator at the other (see Fig. 30 and discussion in PART I). The small height of the resonator system from pole piece to pole piece makes it necessary to use a loaded line, in this case of H-shape cross section. The nature of the output circuit may be seen in the photograph of a cutaway model of the 4J52 in Fig. 75.

The transformation from the wave guide impedance of about 400 ohms through the iris formed by the junction of the H-shape and rectangular wave guides and through the $\lambda/4$ section inserts a resistance of about 2 ohms in series with the resonator to which the output circuit is connected. End effects make the desired transformer length differ slightly from $\lambda/4$. The length necessary to give a pure resistance at the input end was determined by measurements in a 10 cm. model with which the distance from the rectangular wave guide to the first voltage minimum in the H-section could be measured.

The vacuum seal in the wave guide output circuit was made at a circular window sealed in a Kovar cup, mounted between choke couplings as seen in Fig. 75. The diameter of the window and its thickness were chosen so that the reflection coefficient of the window would be quite low over a broad band of wavelengths near 3 cm. Insertion of a dielectric such as the glass window into the wave guide line increases the capacitance per unit length of the line. Compensation for this to bring the characteristic impedance back to the normal value is done by increasing the inductance per unit length over the same region. This may be accomplished by reduction of the long dimension of the wave guide resulting in a nearly square cross section. The circular opening, preferable for glass sealing, is a compromise, the critical diameter being determined by experiment. The relatively large size of the window makes it capable of withstanding RF voltage breakdown even at very high power.

Control of the output coupling is most readily effected by varying the width of the slot in the H-section. The over-all transformation properties of the H-section have been analyzed theoretically and the results confirmed by measurements on output circuit models. By means of this analysis it is possible to estimate fairly well the effect of changes in the slot width upon the pulling figure. To hold the latter within production limits of $\pm$ 2 mc/s, the slot must be held to ±0.001 in.

The output circuit of the magnetron is formed by building up the assembly from sections in what has been called a "sandwich" type construction. The resonator system and the central portion of the output H-section are machined into one slab of copper. The "sides" of the H-section are

milled into this piece on either end. This center slab is "capped" on each end by a slab of copper, in which are contained the magnet pole pieces and appropriate surfaces on which to build the rest of the magnetron. Some of the details of this structure may be seen in the cutaway model of Fig. 75.

Despite the large impedance transformations involved, the output circuit is quite frequency insensitive. As it appeared that these magnetrons might be made tunable, considerable attention was paid to this characteristic. Numerous tests of the output circuit and its component parts were
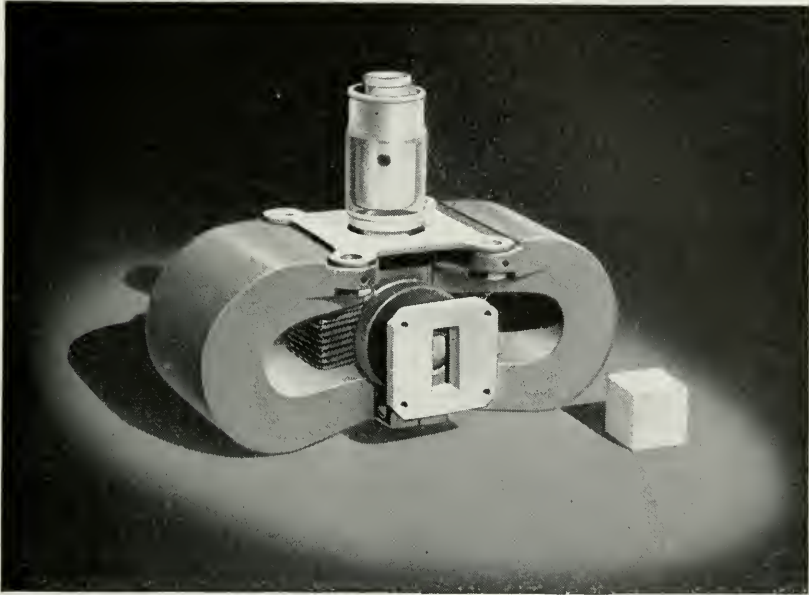
Fig. 73—An external view of the 4J50 "packaged" magnetron (280 kw., 9375 mc/s). Note the circular glass window in the wave guide output circuit and rugged axial input lead which requires no external protecting boot.

made over the 8500–9600 mc/s band. The iris and quarter wave length guide section were studied theoretically in this regard as well. The transformer properties of the window and choke coupling combination is the most frequency sensitive part of the entire output circuit. By adjustment of the distance from the H-section to the window, it was found possible to cancel some of the sensitivity of the two parts.

External views of the 4J50 (4J78) and the 4J52 magnetrons are shown in Figs. 73 and 74 respectively. The internal view of the 4J52 is shown in Fig. 75.

Except for magnet size and cathode supporting structure all three of

these magnetrons are essentially identical.   Geometrical and performance data are given in Table III.   As may be seen, these magnetrons represent an increase in power capabilities by a sizable factor over the 725A.   The 4J52 magnetron has been used extensively under the severe conditions of 5 μs pulse duration, a repetition rate of 200 pulses per second, 15 kv. and 15 amps. input.   Under these conditions it has not been possible to eliminate entirely a tendency to arc, although good performance is obtained.
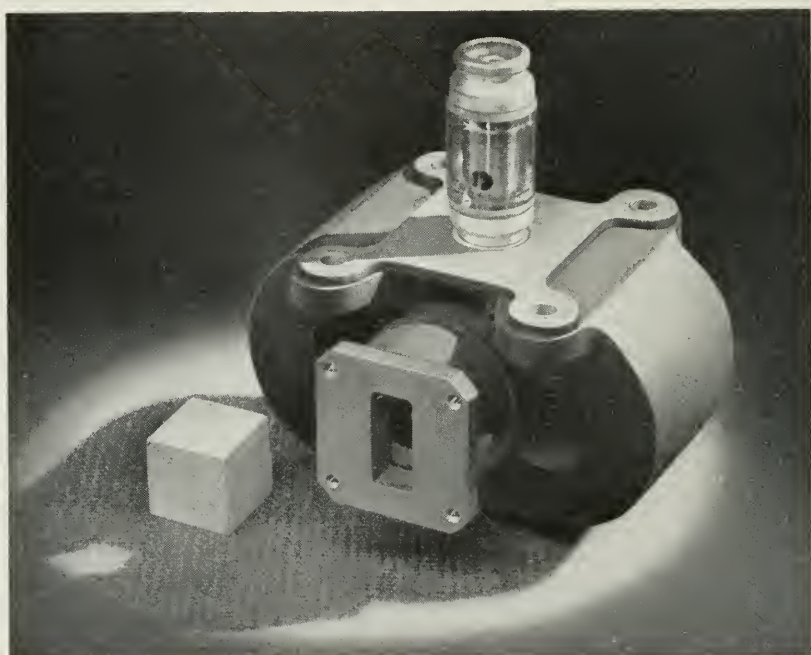
Fig. 74—The 4J52 "packaged" magnetron (100 kw., 9375 mc/s).

In most respects these magnetrons represent the optimum achieved in magnetron design for wavelengths near 3 cm. during the war.   Largely because of the lack of hampering electrical and mechanical restrictions, it was possible to make use of all the latest and best techniques in the design of each of its parts.   Thus each possesses the desirable features of "packaging," axial cathode mount, wave guide output, and high efficiency at low pulling figure.

20. Magnetrons for Wavelengths near 1 Centimeter

20.1 *Preliminary Work:* Intensive work on magnetrons for wavelengths near 1 cm. began when a joint program with the Radiation Laboratory at

Columbia University was undertaken to prepare for manufacture a variation of a magnetron developed there. Prior to this, some work primarily of an exploratory nature had been done in our Laboratories. When it had been demonstrated that scaling from the 10 cm. range yielded an efficient 3 cm. magnetron, similar attempts were made in scaling to wavelengths of 1 to 2 cm. Magnetrons of such wavelengths, obtained by scaling from eight
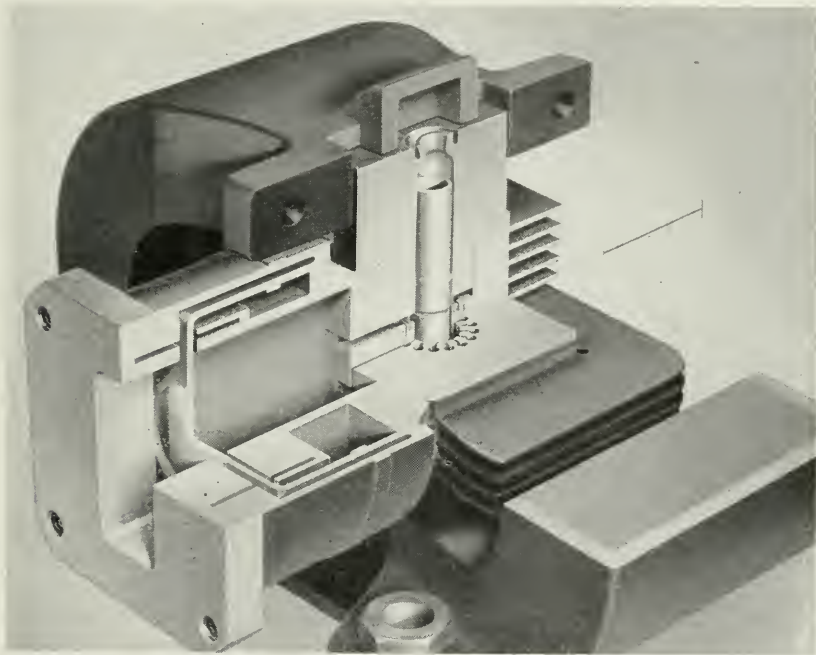
Fig. 75—A view of a cut-away 4J52 "packaged" magnetron (100 kw., 9375 mc/s) which in internal details is essentially like the 4J50 magnetron (280 kw., 9375 mc/s). Of special interest are: the axial cathode construction including permendur end pieces and radiative extension; the quarter wavelength transformer of H-shape cross section (between the output resonator and the line wave guide); and the wave guide window construction with associated chokes.

resonator, 10 cm. models like the 706A-C, were unstrapped while others were strapped with the early British type of strapping and with double ring type strapping. Others were made by scaling from twelve resonator, 3 cm. models and were strapped with both double and single ring straps. The output circuits used were of the loop and coaxial type.

Although some of these magnetrons were faulty, primarily in their output circuits, the majority oscillated. One magnetron of the unstrapped eight resonator variety, the most successful type, was operated over a consider-

able period in our Laboratories. Although no magnetron was developed for manufacture from this work, the experience gained was very useful in the concurrent work at 3 cm. where a similar development program of small magnetrons was under way. Specific developments in this category which should be mentioned are: the use of inserts in which the magnetron resonator system is machined, separable from the rest of the magnetron structure; the use of a double lead cathode input seal to reduce over-all thickness; assembly tools for cathode and strap alignment; and the work with small oxide coated cathodes. The work was discontinued here because of extensive commitments for work at 3 cm. and because an active and successful program of development of magnetrons near 1 cm. wavelength was under way at Columbia University.

The work at the Columbia Radiation Laboratory from the start had been concentrated on magnetrons of very short wavelength. The first Columbia magnetron that was reasonably successful was unstrapped, having a vane type resonator system. It was manufactured in small quantities as the 3J30. Later, the "rising sun" resonator system was discovered there 3J30. Later, the "rising sun" resonator system was discovered as a means of obtaining mode frequency separation. The small scale manufacture was then shifted to the new anode structure, the new magnetron being the 3J31. It was a satisfactory magnetron operating at about 14 kv. and 14 amps., in a magnetic field of 7200 gauss, at a pulling figure of approximately 25 mc/s, and a power output of about 35 kw. The magnetic field was obtained by an external, separable magnet. The cathode was supported by the conventional radial leads.

In addition to the work on the "rising sun" structure, work had continued at the Columbia Laboratory on strapped resonator systems for use at wavelengths near 1 cm. Performance of these magnetrons was quite comparable to that of the "rising sun" variety.

20.2. *The 3J21 Magnetron:* The Bell Laboratories undertook, in collaboration with the Columbia Laboratory, to design a "packaged" version of the 3J31 magnetron for manufacture by the Western Electric Co. At the time, it had not been decided whether the new magnetron should have a strapped or a "rising sun" resonator system. Considerations having primarily to do with manufacturing techniques indicated the latter to be preferable although the former would have been possible. Accordingly, work was started on a "packaged", "rising sun" magnetron to oscillate at 1.25 cm. wavelength (24,000 mc/s). It was to have wave guide output like the Columbia model and axial cathode mount dictated by the "packaged" construction. Operating conditions were to be 15 kv., 15 amps., a pulling figure of 25 mc/s, and as much peak power as could be obtained. It was coded the 3J21.

In the development of the 3J21, the "rising sun" resonator system used by the Columbia Laboratory was adopted practically without change. The ratio of the natural frequencies of the large and small resonators is approximately 1.8. The anode diameter is 0.160 in. and the anode length 0.190 in. The structure is fabricated by the so-called "hubbing" technique,[26] which had been brought to a high state of development at the Columbia Laboratory for the purpose. In this technique, a hardened steel die or hub, machined to be the "negative" of the desired contour, is forced by high hydraulic pressure (of order 250,000 lb./sq. in.) into a copper blank. The hubbed blank, after trimming and turning to size, becomes the resonator system and body of the magnetron. The proper contour to receive the wave guide output circuit is then bored.

The cathode of the 3J21 magnetron has a diameter of 0.096 in. and a length of active coating of 0.165 in. At 15 amps. peak current, the current density at the cathode surface is about 50 amps./sq. cm.—a considerably higher value than that in magnetrons of longer wavelength. Furthermore, the back bombardment of the cathode in this magnetron is about 10 per cent of the input power as compared to the 3 to 6 per cent in longer wavelength magnetrons. For these reasons, the cathode was one of the severest problems of the whole design.

The first axially mounted cathodes were supported on Kovar tubes sealed to the input glass structure. The heater lead was carried down the center of this tube through a glass bead in the input end. The structure resembled that of the 2J51 seen in Fig. 71. Since the hole through the magnetic pole piece was initially 0.100 in. and the cathode end disks 0.130 in. in diameter, the cathode was assembled onto its support after the Kovar tubing was in place. However, in quantity production the cathode could better be assembled as a unit before attachment to the outer glass and pole piece structures. At the expense of magnetic field strength and uniformity, the hole in the pole piece was enlarged to accommodate the entire cathode structure. No ill effects were observed and no special features such as the permendur cathode end structures of the 4J50 magnetron were found necessary.

The problem of the dissipation of the considerable heat of back bombardment was complicated by the necessity of activating the cathode at a temperature of 1050°C. The requirements of heat dissipation and activation oppose one another, one calling for low thermal impedance of the cathode

---

[26] The technique is here called "hubbing" rather than hobbing for two reasons: The term hobbing has a meaning in machine practice quite apart from the technique in question. The term hubbing is used for processes in which the interior of a piece is removed by forcing a member into it. It presumably arose from the early practice of making axle holes by forcing a hardened steel member through the wrought iron wheel hub. Furthermore, it has an analogous usage in coining.

support, the other, high. To meet each requirement satisfactorily with the first cathode structure was found very difficult if not impossible.

A major step forward was taken in the design of the so-called "soldering iron" cathode in which the heater element is placed in a larger part of the cathode lead, heat being conducted to the cathode surface through a solid rod. The cathode itself is solid, making for greater rigidity and lower thermal impedance. The heater element is placed where it may be made considerably larger and more rugged than if placed inside the cathode, whose diameter is 0.096 in. A view of the 3J21 magnetron cathode structure is to be seen in Fig. 78. The heater is contained within the section of larger diameter immediately adjacent to the cathode.

The "soldering iron" cathode presented more difficult problems of heat conduction and dissipation than the older design. For an operating temperature of 800°C at the cathode surface, it is necessary to heat the heater chamber to about 1000°C, whereas the necessary activation temperature of 1050°C requires the temperature of the heater chamber to be about 1300°C. As a result, rather careful design to provide the proper balance between heat losses by conduction and radiation was necessary. Radiation losses, which are the more important type, are increased by extending the cathode rod considerably beyond the active surface, as seen in Fig. 78. The cathode should also have good thermal conducting properties in this extension, throughout the main cathode body, heater chamber, and support. Under normal operating conditions the heater is turned off, cathode heat being supplied solely by back bombardment.

It is apparent that a cathode design of this type calls for a careful choice of materials. They must be highly refractory as well as of good thermal conductivity and structural strength. Copper, silver, and nickel do not meet all of these requirements. The first cathodes of promise were turned from solid stock of so-called machinable molybdenum, complete with cathode end disks and heater chamber. The heater end was brazed to a Kovar detail, subsequently welded to the support cone of the type developed for the 4J52 magnetron. However, molybdenum machines poorly. The tolerances on size did not permit of large scale production of precision molybdenum parts. Consequently it was proposed to make the cathode of tungsten rod which could be ground to shape even on a production basis. The cathode end disks were to be punched or turned from molybdenum or Kovar and brazed to the tungsten rod. The tungsten rod extends into a hollow molybdenum heater chamber, the dimensions of which are not critical. Inside the heater chamber the heater coil encloses the protruding tungsten rod so that better heat conductivity from the heater to the cathode is provided. The heater is brazed to the cathode at one end in the braze between the tungsten and molybdenum. The other end is spot welded to

a heavy central support lead inside the Kovar cone through a hole provided in the cone. The input end of the heater chamber is flared to accommodate the Kovar support cone to which it is brazed. All of the materials are thus of good structural strength. The tungsten and molybdenum are good thermal conductors, the Kovar not. Some parts of the cathode were grit blasted to increase their radiative emissivity.

The total axial motion of the cathode by thermal expansion from cold start to operation is 0.008 in. The cathode is offset axially by this amount when installed.

The cathode surface consists of a nickel matrix base, like that developed for the 725A magnetron, into which the active coating is impregnated. Attempts were made to increase the thermal conductivity through the 0.008 in. thickness of matrix. Larger particle size was used. Ball milling the nickel powder before sintering resulted in more dense particles. A later improvement was effected by making the matrix oversize by a mil or two, impregnating it with the active coating, and then compressing it to size in a mold. This procedure also resulted in a surface far superior in smoothness and regularity to that previously obtained.

The first output circuit used in the 3J21 magnetron, like that in the Columbia model, involved a quarter wave length transforming section of low characteristic impedance (about 20 ohms), extending from an iris in the "back" of one of the large resonators directly into the output wave guide. This section was 0.350 in. high and about 0.0125 in. wide. It was made by milling a slot into solid bar stock, the top of which was closed by brazing on a copper plate. A quarter wavelength section of this line was brazed between the anode body and the output wave guide piece which carried the choke joints and wave guide window seal. This latter piece was fabricated by hubbing a rectangular wave guide in a copper cylinder into which the circular choke was later turned. The wave guide window consisted of a circular glass disk sealed into a Kovar cup like that used in the 4J50 magnetron (compare Figs. 77 and 75). In the 3J21, as in the 4J50, the external choke facing the wave guide window and a short section of wave guide were later incorporated as part of the magnetron.

Holding the pulling figure of the operating magnetron to a specified value presented a serious problem. The problem was one of securing uniformity of the narrow transformer section and, to some extent, of the resonator to which it is attached. A variation of 1 mil in the width of this transformer produced 3 mc/s change in the pulling figure. Even with an improved mechanical design, the spread in output characteristics was uncomfortably large. Part of the difficulty resulted from the formation of solder fillets in the transformer section during brazing, the size of which could not easily be controlled. A rigid inspection of both electrical and mechanical

properties of each magnetron before sealing and pumping was an absolute necessity to insure reproducibility. The use of an iris coupling between the resonator and the wave guide presented a possible solution. However, the iris required would be too large to be placed directly at the back of the resonator. It was proposed to put a hubbed rectangular resonant iris at the back of the resonator and a decoupling iris a half wave length distant in the wave guide. This construction provided a resonant cavity between the two irises in which the storage of energy would provide some degree of frequency stabilization. The structure of this output circuit may be seen in the photograph of a cutaway model in Fig. 77. Both higher efficiency of

Fig. 76—An external view of the 3J21 "packaged" magnetron (60 kw., 24,000 mc/s).

operation and lower pulling figure than obtained with the earlier circuits resulted. As average values, an increase in over-all efficiency from 24 to 28 per cent and a drop in pulling figure from 25 to 18 mc/s may be cited.

The RF voltage breakdown strength of the wave guide window was marginal. A new design developed at the Columbia and M. I. T. Laboratories, incorporating a larger window and "streamlined" wave guide contours adjacent to it, was modified slightly and used in the 3J21.

The magnetic pole gap was made 0.290 in. This was as small as was felt practical. This and other steps were taken in an effort to save as much as possible on magnet weight.

The mechanical problems connected with the fabrication of the 3J21

magnetron resulted largely from the small tolerances it was found necessary to impose. The tolerance allowed on radial location of the cathode was ±0.002 in.; on transformer slot width (when this output was used), ±0.0005 in.; and on dimensions of the anode structure, ±0.001 in. Deformations, such as those produced by placing a magnetron with slightly misaligned pole pieces on a strong magnet, had to be carefully avoided. The importance
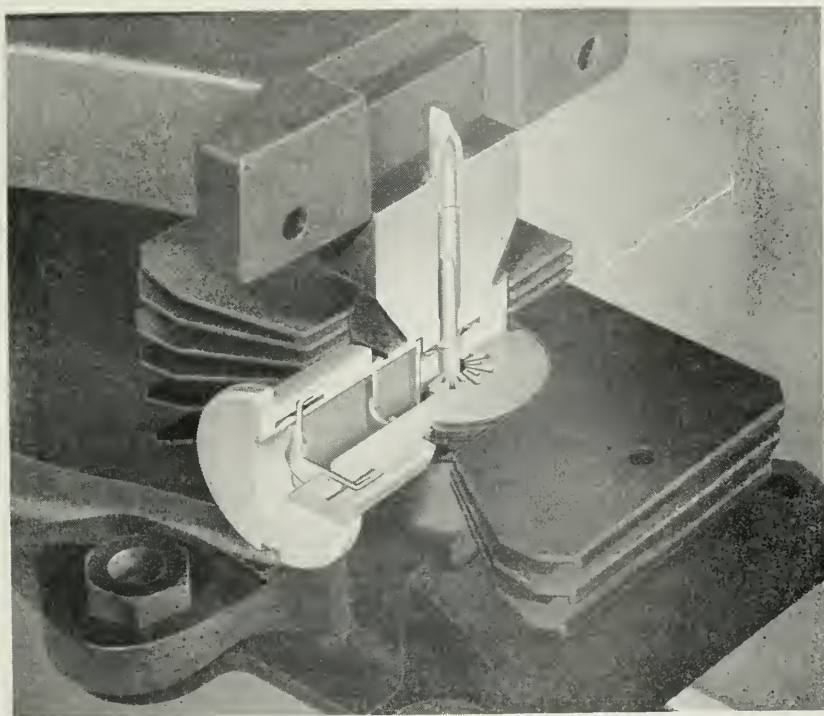
Fig. 77—An internal view of the 3J21 "packaged" magnetron (60 kw., 24,000 mc/s). Note the "rising sun" resonator system, the stabilizing cavity in the wave guide output between a rectangular resonant iris at the "back" of the output resonator and a circular decoupling iris a half wavelength distant, and the beveling of the wave guide edges adjacent to the wave guide window.

of cathode alignment, for example, may be seen from the fact that across the anode-cathode space, which in the 3J21 is only 0.032 in., there is an operating voltage gradient of 165 kv./in. This gradient is doubled in the region where the cathode support passes through the hole in the pole piece, at which point there is a nominal clearance of 0.015 in. Thus the elimination of small burrs and small radii was essential.

An external view of the 3J21 magnetron is shown in Fig. 76, an internal view in Fig. 77. Operational and other data are to be found in TABLE IV.

The principal operational problems encountered in the 3J21 magnetron were the occurrence of so-called "gassy tubes" and some difficulties with "moding." Despite numerous precautions taken during fabrication, gas, identified spectroscopically as hydrogen, was evolved during operation in a discouragingly large number of magnetrons during the developmental work. By the use of a zirconium getter attached to the cathode support cone the difficulty could be circumvented. Also it had been found possible

TABLE IV
THE 3J21 PACKAGED MAGNETRON AT 1.25 CENTIMETERS

N	18	
r_c (in.)	0.048	
r_a (in.)	0.080	
h (in.)	0.186	
Magnet gap (in.)	0.290	
Weight (lb.)	6.6	
Resonator system	"rising sun"	
Resonators	vane type	
Resonator λ ratio	1.8	
λ (cm.)	1.25	
f (mc/s)	23986 ± 240	
Nearest modes	$n = 4, n = 8$	
λ separation (%)	$+25, \quad -10$	
Q_0	1400	
Q_{ext}	580	
η_c (%)	70	
Output circuit	wave guide with λ/2 stabilizing cavity	
Stabilization factor	2	
V (kv.)	15.0	15.0
I (amps.)	15	15
B (gauss)	8000	8000
τ (μs)	0.5	0.25
pps	1000	2000
P_0 (kw.)	60	60
η (%)	26	26
η_e (%)	37	37
PF (mc/s)	17	17

to "clean up" a gassy tube temporarily by running the cathode very hot. This led to a pumping cycle which greatly reduced the occurrence of hydrogen.

Considerable effort was expended in attempting to track down the source of the hydrogen evolution. The iron pole pieces were found to be the major offender, but some hydrogen was evolved from other parts including the cathode. Cleaning and plating solutions release atomic hydrogen which readily permeates the iron pole pieces and other parts. Various

methods of treating the parts, such as preglowing in vacuum, were used. These largely eliminated the difficulty.

An interesting feature of the problem was the relatively infrequent occurrence of hydrogen in magnetrons built at the Western Electric Co. The reason for this was finally traced to the precaution, prescribed by the safety engineers, that the sintering of the copper plate on the pole pieces be done in an atmosphere of 95 per cent nitrogen and only 5 per cent hydrogen. All sintering and brazing operations in our Laboratories had been done in a 100 per cent hydrogen atmosphere. Pole pieces sintered in the predominantly nitrogen atmosphere were found to be nearly as free of hydrogen as the vacuum preglowed parts. However, the use of the zirconium getter was continued in models built at the Laboratories as added insurance against recurrence of the difficulty.

The other operational fault of the 3J21 magnetron was a tendency to oscillate in a mode other than the π mode under certain circumstances. Examination of a large number of magnetrons, opened after operation, showed that the only observable causes were resonator distortion, the presence of brazing flux or other foreign material, and off-center cathode location. As in other magnetrons, the tendency to "mode" was greater when heavily loaded or when operated with a sharply rising voltage pulse. Western Electric engineers concluded, as the result of an extended study of magnetrons in which the cathode position was changed during operation by flexing the input lead, that a cathode position 2 to 3 mils off center toward the output resonator was desirable. A tool was developed with which a permanent change in cathode location could be made on operating magnetrons. By its use many "mody" magnetrons were reclaimed.

An interesting phenomenon, discovered during the development of the 3J21 magnetron, is that known as the "cold start." Whereas most magnetrons may be started with difficulty with a cold cathode, the 3J21 was found to start readily under these conditions over several discrete bands of magnetic field intensity. Although the phenomenon is not well understood, some correlation with the work of Posthumus[13] has been made.

In making tests on the 3J21 magnetron at a duty cycle of 0.001 and a pulse duration of 0.1 μs, the cathode was found to operate at a considerably higher temperature than at 0.5 μs pulse duration and the same duty cycle and average input power. This behavior has since been observed in the 725A and 4J52 magnetrons. It is thought to result from unfavorable conditions during the periods of rise and fall of the voltage pulse, which become a greater percentage of the total pulse duration as the pulse duration is reduced. The effect presumably would not appear if the pulse shape were truly rectangular. With present modulators, however, it

presents a lower limit on pulse duration which may be employed and further complicates the cathode cooling problem.

21. Magnetron Cathodes

From the start of its work with centimeter wave magnetrons, the Bell Laboratories has been engaged in an extensive program of magnetron cathode testing and design. The range of cathode size is illustrated in Fig. 78. The largest cathode, shown at the extreme left, is that of the 5J26 magnetron, operating at about 23 cm., the smallest, that of the 3J21 magnetron operating near 1 cm. The range of operating conditions to be met, as well as the magnitude of the problem at shorter wavelengths, is illustrated by these two extremes. The cathode of the 3J21 magnetron, although its surface area is only 0.31 sq. cm., is called upon to deliver peak currents comparable to those demanded of the cathode of the 5J26, having a surface area of 17.1 sq. cm. The ratio of current densities is 25 to 1. The peak back bombardment of the two cathodes is comparable. Other cathodes shown in Fig. 78 operate under conditions intermediate between these extremes. Needless to say, the 5J26 cathode is operated quite conservatively, the 3J21 cathode under extremely severe conditions.

Little difficulty with cathodes has been experienced in pulsed magnetrons above 10 cm. wavelength. For the most part they are plain, nickel sleeves coated with active material. The highest emission density (10.1 amps./cm.²) used successfully with this type of cathode was in the 720A-E, which operated satisfactorily at 1 μs but was not satisfactory at 5 μs. The operation at 5 μs required a modification of the cathode as described. When the simple sprayed cathode was tried in developmental models of the 725A magnetron, the life at 1 μs was of order 10 hours. At the high emission density required (approximately 30 amps./cm.²) the active coating was rapidly lost as a result of arcs, and frequently the nickel support itself was fused and vaporized.

The problem of arcing to the cathode surface in small magnetrons was soon recognized as the most important cathode problem. It prevented steady operation and hastened destruction of the cathode surface and the end of useful magnetron life. All of these cathodes were found to require an initial break-in period during which the arcing is particularly violent. As the input voltage and current are gradually increased, the violent arcing gradually subsides. The tendency to arc at any time subsequent to this initial period depends on the nature of the cathode surface, but beyond that, it depends also on the operating conditions to which the magnetron is subjected. Increase in power input, either as increased voltage or current, or both, rapidly increases the frequency of arcs. Similarly, increased pulse
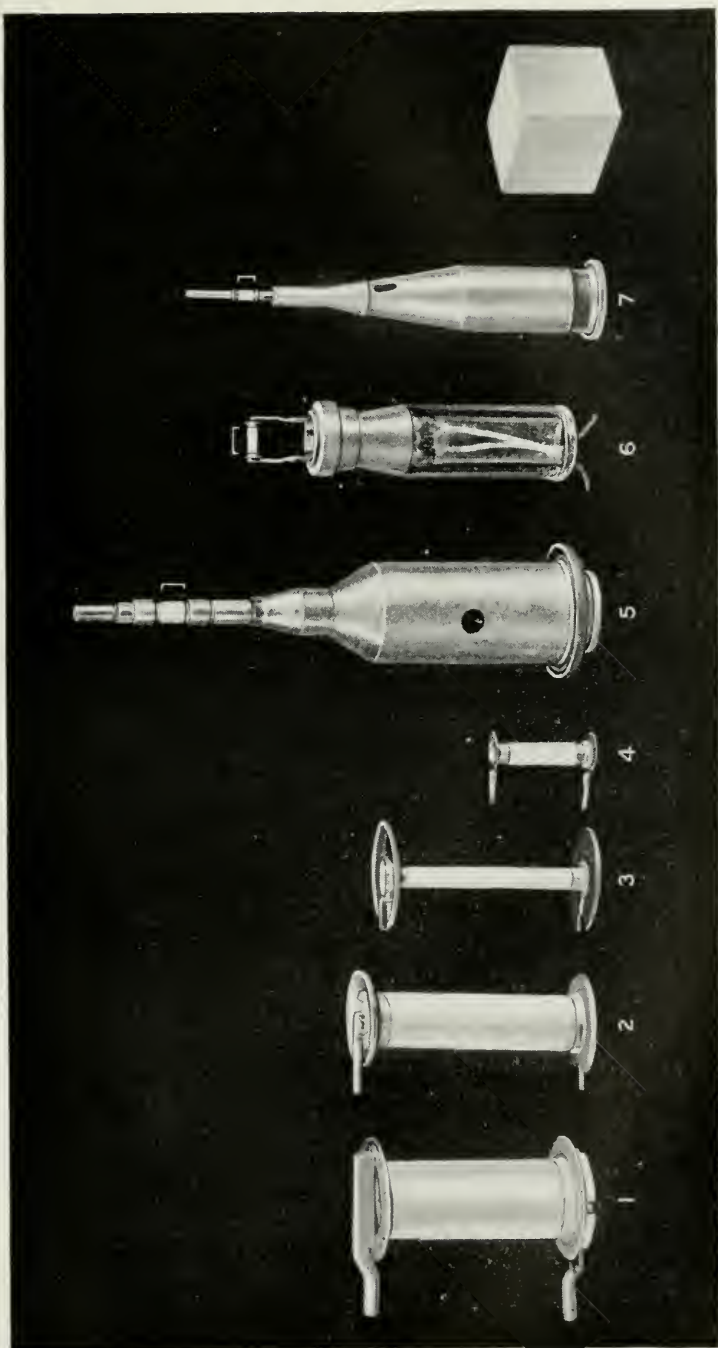
Fig. 78—A group of cathode structures for centimeter wave magnetrons. From left to right they are: (1) the 5J26 cathode—oxide coating on roughened nickel base; (2) the 4J21-30 cathode—oxide coating on roughened nickel base; (3) the 720A-E cathode—metallized, oxide coating on nickel wire mesh base; (4) the 706AY-GY cathode—oxide coating on smooth nickel base; (5) the 4J50 cathode and support structure—oxide coating on sintered, nickel powder, matrix base; (6) the 725A cathode and support leads—oxide coating on sintered, nickel powder, matrix base; and (7) the 3J21 cathode and support structure—oxide coating on sintered, nickel powder, matrix base. The active cathode portion in each of the structures (5), (6), and (7) is indicated by a bracket.

length makes the arcing much worse; in the 725A for example, changing the pulse length from 1 μs to 2 μs may increase the arcing rate by a factor of 10 to 20. A decrease in recurrence rates with some types of cathodes also causes a noticeable increase in the rate of arcing.

In the great majority of cases, magnetron cathodes have employed coatings of the ordinary strontium and barium carbonates as the source of primary and secondary electrons. All-metal, secondary emitting cathodes with primary emitting, starter cathodes have been tried in some laboratories with some success but have not yet come into general use. The efforts at the Bell Laboratories have been directed toward developing and improving the oxide coated cathode along lines making it more nearly possible to satisfy requirements (3), (4), (5), and (6) listed in Section 10.7 *Magnetron Cathodes* of PART I without impairing the ability to meet requirements (1) and (2).

Most of the cathode developments were made in connection with the 725A. This magnetron served as a convenient "laboratory" or "proving ground" for magnetron cathode studies. Its cathode is small enough and the demands made upon it stringent enough to make apparent any cathode weaknesses and any improvements which may be made. A large number of life racks were kept in constant operation for 725A magnetron life tests over a long period of time under various conditions in which the cathodes were generally run to destruction.

Attempts were made to make the data reproducible by controlling the activation, by controlling the initial break-in of the operating magnetron, and by devising means of quantitatively determining some measure of the adequacy of any given cathode. The last of these items involved the development of an automatic counter which registers the number of arcs or bursts of arcs which cause the current to exceed a predetermined value. By recording the accumulated number of arcs at intervals throughout the life of the magnetron, it is possible to get a picture of the arcing pattern with life. Such counters have done much to put the life testing and initial break-in of magnetrons on a semi-quantitative basis. The smoothed curves shown in Fig. 80 were obtained through the use of these counters.

Early attempts at building a cathode upon which sufficient active material may be held, made at both the M. I. T. Radiation Laboratory and the Bell Laboratories, involved winding coated cathodes with nickel wire to provide a reservoir of active material under the wires. In this manner, the material was protected from direct arcing effects, allowing the active material to migrate slowly to the outside surface of the wires. This general idea was greatly improved upon when the wire winding was replaced by a woven nickel mesh into the interstices of which the active material was packed. The

nickel mesh, being spot welded or sintered directly to the nickel base, formed a part of the cathode foundation, helping to reduce the coating resistance. It was found that arcs, rather than striking directly to the active material, tended to strike to the nickel wires. In spite of this fact, sudden bursts of current in arcing would erupt some of the active material from the cathode.

The mesh used in the 725A cathode is made of 6 mil nickel wire woven with 75 wires to the inch. The radial thickness of the mesh on the cathode is 10 mils. The average amount of active double carbonate mixture which the cathode holds is 23 mg. per sq. cm. The mesh cathode is shown in Fig. 79 in three stages of its fabrication. With this cathode, it was possible to
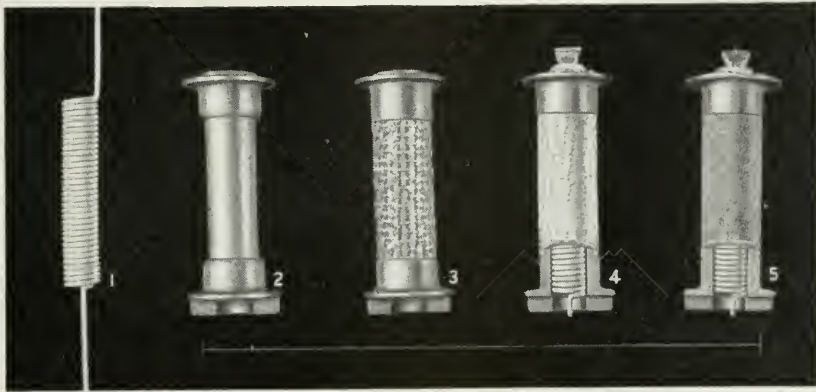
Fig. 79—725A cathode construction. Shown are: (1) the heater element; (2) the machined nickel cathode blank; (3) the cathode blank with the nickel wire mesh, welded or sintered in place; (4) the mesh type cathode with oxide coating applied, cut away to show heater element in place; (5) the more recent cathode with sintered, nickel powder, matrix base (uncoated).

produce 725A magnetrons having a guaranteed life at rated operating conditions of at least 500 hours, the average value being considerably higher. The arcing characteristic of the mesh cathode, excluding the initial break-in period, is shown in Fig. 80. The abrupt failure is typical of magnetron cathodes. At shorter pulse lengths failure occurs considerably later but in the same abrupt fashion. The life curve of the early type of cathode consisting of an oxide coated nickel cylinder, if plotted with the data on other cathodes in Fig. 80, would be crowded very close to the axis of zero hours.

Even with cathodes of mesh construction, operation of the magnetron was generally inadequate at 5 μs pulse duration. Furthermore, considerable break-in time was necessary before stable operation at 1 μs pulse duration was achieved. Some of the first mesh cathodes used in the 725A,

for example, took as long as 45 minutes of intense arcing during gradual increase of input voltage and current before satisfactory operation was attained. It seemed unlikely that the arcing during the break-in period could be completely eliminated, but it was found that improvements in cathode construction which made steady operation at higher current possible also reduced the time of initial break-in to attain these conditions.

It appeared that the resistance of the cathode coating might be lowered by distributing nickel more uniformly throughout the coating. Accord-
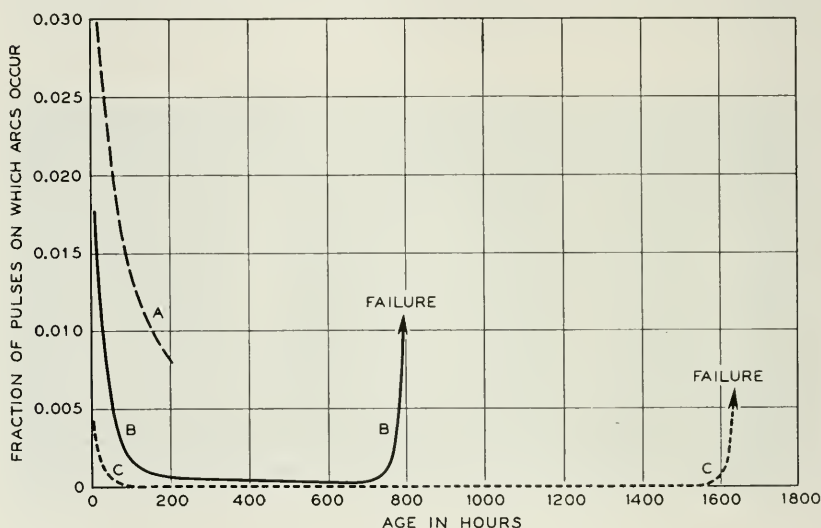
Fig. 80—Curves of arcing during life for 725A magnetrons, having various types of cathodes, operating at 5.7μs pulse duration, 165 pulses per second, and 210 kw. peak power input. It is to be emphasized that these conditions are considerably more stringent than the normal operating conditions. Curve (a) is for a cathode having oxide coating on a nickel wire mesh base; curve (b), metallized coating on a nickel wire mesh base; curve (c), either plain oxide or metallized coating on a sintered, nickel powder, matrix base.

ingly, mesh cathodes were coated with a mixture of double carbonates in which was distributed a fine nickel powder of high purity having an average particle size of 2 microns. The carbonate particle size is between 1 and 2 microns. The amount, found not to be critical, was 55 per cent by weight of the combined dry ingredients of the coating mixture. Although this reduced the amount of active material present on the cathode surface, as seen in Fig. 80 it resulted in a cathode having considerably improved arcing characteristics with considerably longer useful life. This cathode construction was adopted finally for the 725A production and was also utilized in the 10 cm. high power magnetrons, the 4J45–47, for operation at 5 μs pulse duration.

The next step in the development was the replacement of the nickel mesh by a sintered matrix of coarse nickel powder (see Fig. 79). The average particle size was 55 microns. The active material is packed into this matrix in much the same manner as in the mesh cathode. Little difference was observed in the addition of fine nickel powder to the active coating in this type of cathode. The matrix cathode has the best life characteristic of any yet devised for the magnetron oscillator. Its life characteristic is shown in Fig. 80 with those of the earlier types.

The sintered matrix type of cathode has been used extensively in the 4J50 and 4J52 magnetrons at 3 cm. and in the 3J21 magnetron at 1 cm. In these cathodes, it was important to have high heat dissipation. This was achieved by designing a high thermal conductivity support and providing considerable radiating area immediately adjacent to the cathode surface by extending the cathode, as shown in Figs. 75 and 77, into the pole piece opposite that carrying the cathode mount.

22. Acknowledgments

Contributors to this Issue

J. B. Fisk, S.B. in Aeronautical Engineering, Massachusetts Institute of Technology, 1931; Redfield Proctor Traveling Fellow, Trinity College, Cambridge, 1932–34; Ph.D. in Physics, M.I.T., 1935; Junior Fellow, Harvard University, 1936–38; Associate Professor of Physics, University of North Carolina, 1938–39; Bell Telephone Laboratories 1939–. Engaged in magnetron work in the Electronics Research Department during the war; now in the Physical Research Department.

Homer D. Hagstrum, B.E.E., 1935; B.A., 1936; M.S., 1939; Ph.D. in Physics, 1940, University of Minnesota; Research and Teaching Assistantship, University of Minnesota, 1936–40; Bell Telephone Laboratories, 1940–; engaged in the development of magnetrons during the war in the Electronics Research Department; now in the Physical Research Department.

Paul L. Hartman, B.S. in E.E., University of Nevada, 1934; Ph.D. in Physics, Cornell University, 1938; Instructor in Physics, Cornell University, 1938–9; Bell Telephone Laboratories 1939–; has been a member of the Electronics Research Department, being engaged during the war in the development of magnetrons; now in the Physical Research Department.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

Some Recent Contributions to Synthetic Rubber Research*

By C. S. FULLER

Introduction

WHEN the war put an end to shipments of natural rubber from the Far East, it became evident that synthetic chemistry would be called upon to fill the gap in our supply of this strategic material. We know now how effectively the emergency was met. In less than three years the production of Buna S type synthetic rubber alone had risen to exceed our total prewar comsumption of natural rubber. Few, however, realize the magnitude of the effort and the extent of the cooperation between groups of experts that was essential for the achievement of this success.

Rubber companies in this country had been experimenting with synthetic substitutes for natural rubber for some time before the present war began. None of these products, however, was sufficiently advanced either from the stand-point of raw materials or in regard to the knowledge of its properties, to warrant production on a large scale as a substitute for natural rubber during the emergency. In 1942, following the advice of the Baruch Committee, we decided to place chief reliance on Buna S, the butadiene-styrene synthetic rubber which the Germans developed about 1934. In addition, considerable support was given to the domestic synthetics, Neoprene, Thiokol and Butyl. The latter rubbers, however, were not considered as useful for tires as Buna S.

Making Buna S in this country and fabricating it were not simple, however. The Germans had kept the details of the process secret and restricted shipments of the product. Besides, as we have since found out, the German chemists did not have any too complete control of the process themselves and the type of rubber made by them, as shown by samples obtained indirectly, was not satisfactory for use on American processing machinery. Our engineers and research men were therefore faced with the problem of setting up a process on an enormous scale to turn out a product which could be used in our tire plants and which would give satisfactory service on the road. Fortunately for us, a few companies had acquired enough knowledge

both from German sources and from their own researches to warrant taking the gamble.

As a part of the large program laid out under the auspices of the Government in 1942, provision was made for a cooperative research and development effort to parallel and to contribute to the constructional program designed to provide the much needed rubber. A large number of company laboratories as well as universities contributed to this research. The Bell Telephone Laboratories because of its past contributions in this field of synthetic polymers was asked to participate in this program. The present discussion is intended to describe part of the Bell Laboratories investigations directed toward the improvement of Buna S type rubber, particularly work relating to the characterization and control of the final copolymer.

In order to present the material in a logical and understandable form to readers unfamiliar with the subject-matter, brief mention will be made of the history of the synthetic rubber problem and of progress in the knowledge of polymeric substances during recent years.

The Problem of Synthetic Rubber

The problem of synthesizing natural rubber is almost as old as man's curiosity about the nature of rubber itself which began when Faraday in 1826 first showed it to be a hydrocarbon having the formula $C_{10}H_{16}$. Experiments done by Williams in 1860, in which he obtained isoprene from natural rubber and by Bouchardat in 1879, who showed that isoprene could be polymerized to a rubber-like material, represent about as close as we have come to synthesizing natural rubber in spite of many subsequent efforts. In 1910 particularly, when the price of natural rubber reached $3 per pound, considerable pressure was exerted to bring about this synthesis. Although the chemist failed in this quest his very failure, analyzed in the light of more recent studies on other polymers as well as rubber, has had its virtues. It has emphasized the importance of chemical structure, that is the precise organization of the atoms composing the rubber molecules (in addition to simply the nature of these atoms) in determining the ultimate properties of a polymer.

Although natural rubber eluded synthesis, the early organic chemical work nevertheless laid the basis for our present synthetic rubber. Curiously, much of this pioneering research on synthetic rubber was done in England with the support of strong proponents of natural rubber. However, Germany and Russia were also active contributors. The United States later achieved fame by bringing forth two of the most promising rubbers yet produced, Neoprene and Butyl. The early foreign synthetics were based on the polymerization of hydrocarbons such as 1-methyl butadiene and 2,3

dimethyl butadiene and butadiene itself. They were undoubtedly "rubbers" of a sort but there could be no question about their inferiority to the natural product. Even today the Russians persist in making their synthetic rubber from butadiene and, although there have been improvements, the polymer is still subject inherently to the same fundamental difficulties of structure that existed when it was first synthesized by Lebedev in 1911.

The deficiencies in the early synthetic rubbers and the difficulty of synthesizing natural rubber were appreciated in Germany where in the period 1935–39 several plants were constructed to manufacture synthetic rubber, including Buna S, on a large scale. By polymerizing together butadiene and styrene instead of butadiene alone they achieved several advantages over previous synthetic rubbers. The fact that the best opinion in this country decided in favor of imitating German Buna S, shows that progress in Germany was indeed substantial. As we have already indicated, however, improvements were necessary in both the German product and process if it was to be satisfactory for our use. The product developed in this country and now being currently produced at the rate of nearly 700,000 tons per year, although prepared from the same starting materials as German Buna S, therefore differs from the latter in many important respects. The name Government Rubber-Styrene, abbreviated GR-S, has been given to this product.

History of the Development of Ideas of Composition and Structure of Polimers

All rubbers, both natural and synthetic, as well as all organic plastics and fibers belong to a class of substances called polymers. We now know that they are constructed of large molecules, in turn built up of simple atomic patterns (repeating units) joined end to end. Surprisingly, it was not until about fifteen years ago that this idea gained general acceptance among chemists. Since that time truly remarkable research progress on polymers has been made. It is not our object to present a full account of this work here. Most of it was carried on independently of its application to the synthetic rubber problem but nevertheless has had a profound effect upon it. A brief review of the growth of the present concepts of natural and synthetic polymers will, however, help to emphasize the significance of the more recent researches on synthetic rubber.

For a long time chemists believed that naturally occurring polymers like natural rubber, cellulose and silk were indefinite chemical compounds in which the arrangement of the atoms was so complex as to defy analysis. As has been mentioned, Faraday had shown in the case of natural rubber that carbon and hydrogen atoms were present in the ratio of 16 hydrogens

for every 10 carbons. It was not until much later that it was postulated that rubber, inasmuch as it had the same hydrogen to carbon ratio as isoprene obtainable from it, was a compound in which many isoprene groups were in some manner combined together. Thus, Harries about 1904 was inclined to regard rubber as a sort of association complex representing a combination of relatively small ring molecules held together by van der Waals' attractions[1]. This same view of polymers as associations of small molecules was also applied to cellulose by well-known carbohydrate chemists both in England and in Germany.

The influence of the contemporary colloid chemists helped to promote this idea. Even the term "micelle", applied by them to soap and other aggregates, which are in fact van der Waal's or ionic associations, was unfortunately adopted to describe the structure of many of the organic polymers. In addition, early x-ray studies on natural polymers, because of a misinterpretation of the diffraction patterns, lent further support to these views. For some reason or other it was not appreciated by workers in the field that the x-ray unit cell did not necessarily mark the boundaries of the organic molecule. Hence, since the unit cells appeared to be small, many erroneously concluded that the molecules were small also. It is to Sponsler and Dore[2], working in this country in 1926 on the x-ray structure of cellulose fibers, that we must give thanks for being the first to realize the incorrectness of the older x-ray deductions and to postulate a long primary valence chain structure for cellulose.

The realization that natural organic polymers really consisted of very long chains of primary valence bound atoms, in the strictly organic chemical sense, came surprisingly slowly. Staudinger in Germany beginning about 1926 was most insistent on this view[3], although others including Meyer and Mark were developing the same conception. As early as 1910 Pickle in England had conceived of such a chain type of molecule for natural rubber but unfortunately did not follow it up. As the idea of molecules of large size grew, it became more and more popular to try to measure them. Also there was much effort given to working out the details of the "crystal structure" of the natural products insofar as they could be regarded as crystalline. Here again was an opportunity for argument which is still going on today: just what do we mean by the term "crystalline" when applied to these substances? The answer seems to be that we have all degrees of organization of the molecules, or more correctly parts of molecules, in polymers from the completely chaotic or amorphous in some to highly ordered or what may be called crystalline arrangement in others. We shall have occasion to come back to this subject in our later discussion.

It was logical that the interest of scientists in the constitution and structure of polymers should be lavished on naturally occurring high polymers

rather than on the synthetic ones. But strangely enough it has been the synthetic polymers which have really led us to a more complete understanding of the natural substances and particularly to the explanation of why polymers have the properties they do.

The early work on synthetic polymers, as we have seen, centered around the constitution of natural rubber and efforts to duplicate it. Soon, however, organic chemists found they could make better products from other dienes than they could from isoprene which seemed to be the progenitor of natural rubber. The approach was necessarily empirical—one of trying out a variety of reaction conditions on the chemical compound to be polymerized and studying the properties of the final product as compared to natural rubber. Nearly always the comparison was disappointing. Following this procedure the Germans and the Russians developed their respective competitors for natural rubber from 1910 to the present time. The organic chemistry of polymerization, the reactions whereby the simple unsaturated compounds join up into longer molecules, was, however, very imperfectly understood in 1910 and still is not clear today.

Perhaps it was for this reason that some organic chemists decided to build large molecules by methods in which they had acquired great confidence in regard to how the atoms come together. Emil Fischer, the first of this group, succeeded in synthesizing a polypeptide molecule of known composition and known organic structure which, although smaller in size than the natural proteins, nevertheless was very large compared to the usual organic molecules. This was in 1906. About 20 years later the matter was again opened up in a more general way by Staudinger and his collaborators who synthesized chains built up of alternate carbon and oxygen atoms, the polyoxymethylenes, and showed how such large molecules could give rise to a pseudo-crystalline type of crystal lattice. Then came the simple and beautiful work of W. H. Carothers and his collaborators beginning in 1928, which led to the development of nylon. These compounds and the linear polyesters, which Carothers had (by improvement of the methods of Vorländer[4] and others) prepared, because they were known to contain long chain molecules of definite structure and composition, were ideal compounds to examine in order to determine what factors were truly responsible for observed polymer behaviors. In this way it was hoped to explain the outstanding toughness, high tensile strength, rubberiness, peculiar softening and flow properties and a host of other characteristics of polymers which make these materials so important in life processes and technology. Researches along these lines have indeed shown that the way the various units are combined and the regularity of the atomic arrangements in the units themselves have a profound effect on properties.

This work has also emphasized the importance of size and linearity of the

chain molecules on polymer behavior. For example, the length of the mole-
cules which are present in a polymer is of critical importance to certain
properties such as mechanical strength. These facts, as well as the necessity
for order in the arrangement of the molecular units along the chains were not
appreciated by the early organic workers. That Carothers realized what
many of the older organic chemists did not realize is indicated by his state-
ment made in 1934 that the problem of physically characterizing polymers
in significant numerical units is of the utmost importance and that it should
receive more attention jointly from physicists and chemists.

Some Physico-Chemical Features of Polymers

We have seen very briefly how the quest for the origin of properties of
rubbers and polymeric substances in general led of necessity to a study of the
intimate details of chain molecule structure on the one hand and a study of
the general characteristics of large molecules on the other. Before taking
up the specific researches on GR-S synthetic rubber, however, it will be
helpful to pursue somewhat further the ideas on the formation and constitu-
tion of polymers.

There are two general chemical processes by which polymer molecules are
formed, namely polymerization and polycondensation. Chemists, at times,
use the first term to represent all processes leading to the formation of large
molecules but it is more convenient to distinguish two processes even though
the difference between them is academic in some cases. In polymerization,
chemical molecules called the monomers, become "activated" either by heat
energy or by means of special chemical compounds. In this state they
spontaneously grow at the expense of their unactivated neighbors until the
growth of the chains is abruptly terminated, either by active chains coming
together or by a transfer of energy to other, often foreign, molecules. The
entire growth reaction for any given chain usually takes but a fraction of a
second for completion. When two or more different monomers capable of
polymerization enter together into the same chain molecule formation the
process is referred to as "copolymerization".

In polycondensation, identical or non-identical molecules react to give
large molecules just as in polymerization. The difference is that in the
former reaction a molecule of water (or other substance) is evolved each time
a new molecule is added to the growing chain system. Also the reaction
resulting in chain growth is step-wise in the sense that each added molecule
follows the same steps in reacting that are followed by any other. No
special type of activation on the end of the growing chain is necessary.
Finally, since there is no activated growth, the phenomena of termination
in the sense used above in connection with polymerization do not exist.

Both kinds of polymers are important technically. Thus polystyrene is a

polymerization type of polymer. Nylon on the other hand is a polycondensation polymer. Buna S type synthetic rubber is a polymerization copolymer because it is formed by polymerizing together styrene and butadiene monomers.

One of the important characteristics about reactions leading to the formation of polymers is that they result not in molecules of the same size but in a statistical distribution or mixture of molecules of various sizes. These molecular weight distributions, as they are called, in special cases can be
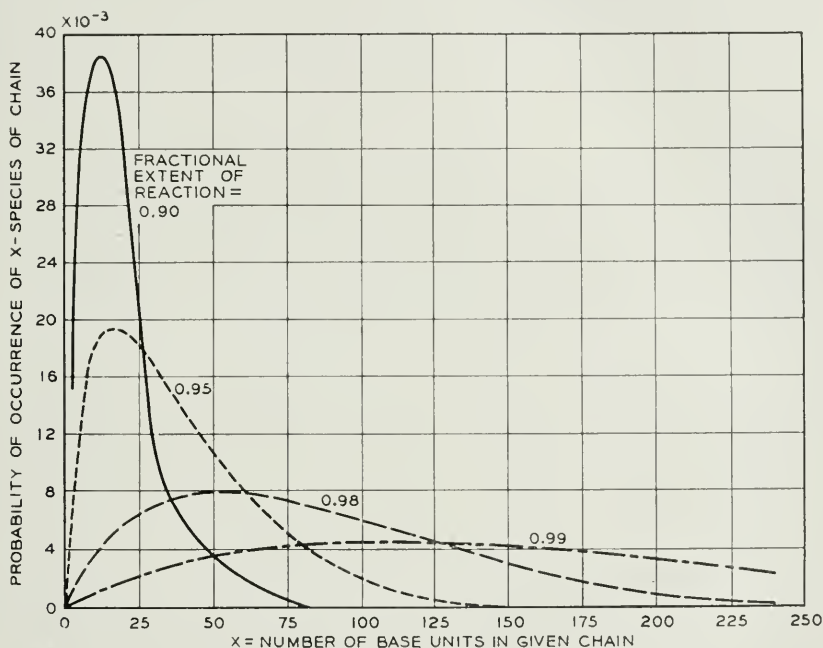
Fig. 1.—Curves showing frequency distribution of molecule species of different chain lengths for linear polyesters (Flory—reference 5).

calculated from the nature of the reaction. In other instances this is not possible, although experimentally it is often possible to arrive at an approximate curve representing a given polymer distribution. Figure 1 shows a series of curves for a linear polyester in which the reaction conditions are such that the calculated curves[5] represent very closely the actual distribution of molecules present. The curves represent the weight fraction of each molecular species present in the mixture at the extent of reaction shown on each curve.

It is customary to speak of an "average molecular weight" therefore in characterizing these polymer mixtures. Several different types of averages

are used for convenience. The two most frequently employed are what are termed the "number average" and the "weight average". If osmotic pressure measurements are made on solutions of the polymer and extrapolated to zero concentration these will lead to a number average molecular weight figure. This average represents what we would obtain if we sorted out the molecules according to molecular weight and counted them. Multiplying each molecular weight (m_1) by the number present (n_1) and dividing by the total number of molecules we obtain the number average molecular weight $(\overline{M}_n)$ or stated mathematically

$$\overline{M}_n = \frac{\Sigma m_i n_i}{\Sigma n_i} = \frac{1}{\Sigma \dfrac{f_i}{m_i}} \tag{1}$$

where f_i is the weight fraction of species of molecular weight, m_i.

Usually the osmotic pressure measurements are difficult to carry out and a simpler measurement, that of dilute solution viscosity (DSV) is performed. This determination consists in measuring the relative viscosity of a solution of the polymer at one given low concentration and calculating

$$(DSV) = \frac{\ln \eta_r}{c} \tag{2}$$

where η_r is relative viscosity and c is the concentration in grams per 100 ml. of solution. A more fundamental quantity usually differing little from the DSV value is the so-called intrinsic viscosity. This is defined as $[\eta] = \dfrac{\ln \eta_r}{c_{c \to o}}$. Measurements are made at several concentrations and extrapolated to zero concentration just as for osmotic pressure. From this value a molecular weight, which may be referred to as a viscosity average molecular weight, can be calculated from the empirical expression $[\eta] = K(\overline{M}_v)^a$ where both K and a are constants over a fairly wide range, and which must be independently determined. In some polymer distributions this viscosity average is very close to the weight average defined by

$$\overline{M}_w = \frac{\Sigma w_i m_i}{\Sigma n_i m_i} = \Sigma f_i m_i \tag{3}$$

where m_i is again the molecular weight of each species and f_i is the weight fraction in which it is present in the mixture. In the example of Fig. 1 the number averages are indicated by the maxima of the various curves. Here the viscosity and weight averages are identical.

In polymers an equally important consideration with molecular size distribution is chain molecule structure. It is convenient to distinguish between micro-chain structure and macro-chain structure. By micro-chain

structure we mean the detailed architecture of the chain molecule over distances of the order of length of the repeating unit. The kind of atoms involved in the unit and their spatial arrangement in regard to atoms in the same chain as well as in the neighboring chains are included in this definition. It is the micro-chain structure which determines entirely the chemical properties of the polymer and to a large extent the physical properties as well. Thus, the influence of solvents, oxidizability, hardness at a given temperature, softening point, ability to crystallize are determined largely by the micro-structure of the polymer.

Macro-chain structure on the other hand refers to the long range form of the chain molecule. It ignores composition and concerns itself with the nature of the molecule as a whole and with its interconnections to other molecules.

Certain terminology has grown up in this connection which can be conveniently defined at this time. We speak of "linear" polymers when primary valence bonds can be traced through the molecules from one end to the other without passing over the same atoms twice. We say "branched" molecules are present when the process of tracing leads us into one or more offshoots from the main chain. When the degree of branching becomes excessive the molecules may become insoluble in good solvents for the linear or slightly branched molecules. When nteworks of molecules are present we say the polymer is "netted" or "cross-linked". In this instance closed paths may be traced and the smaller the paths, the "tighter" or more "intense" is the netting. Netted chain molecule systems are invariably insoluble. Insoluble polymers whether because of intense branching or netting are called "gel". We speak of micro-gel when the gel particles (molecules) are microscopic or smaller in size (say less than 1μ) and of macro-gel when the particles are large.[6] Usually macro-gel as well as the micro-gel is associated with soluble molecule species. These latter are referred to as "sol" and represent the linear or the less branched molecular components of the mixture. The complete description of every molecule present in a polymer mixture is thus a very difficult if not impossible task. We are thus forced to employ a statistical treatment.

In the case of copolymers, still other considerations arise. There is the probability that the reacting components will not react with one another at the same rates they do with themselves. When this occurs the composition of the molecules in the mixture varies, some containing more of one component than others do. Also the order in which the components are arranged along the chains may vary molecule to molecule. Such circumstances of course give rise to varying properties in the copolymer mixture. We shall have occasion to consider these questions below in connection with the development of GR-S.

Early Status of GR-S Synthetic Rubber

The process by which GR-S type synthetic rubber is made is known as the emulsion polymerization process. In it, butadiene and styrene in the proper proportions are emulsified in water with small amounts of catalysts and substances called modifiers which serve to control the plasticity of the polymer. During the reaction period of from ten to twenty hours about three-fourths of the butadiene and styrene are converted into the synthetic rubber. The reaction occurs in such a way that very minute particles are formed and the resulting synthetic latex is suggestive of natural latex. To obtain the rubber itself the latex is coagulated with acid and sodium chloride or with aluminum sulfate and the coagulum washed. After drying the rubber crumbs are baled and shipped to the fabricating factories. The above brief sketch of course does not provide an idea of the many complexities which arise in practice nor of the many process variations which can be used to control the final properties of the rubber. A complete treatment of this subject falls outside the scope of this paper.

When the Baruch Committee advised "bulling through" the synthetic program on the basis of Buna S type rubber, it fixed the chemical composition of the product to a very great extent. We knew then, or shortly afterward, that we would be required to use approximately 690,000 tons of butadiene and 197,500 tons of styrene per year to produce the coplymer rubber. Whatever other components might be employed would be available in only insignificant quantities by comparison. One element of choice remained as far as chemical composition was concerned, namely the proportions in which the two components might be used. German Buna S is supposed to consist of 75 parts by weight of butadiene to 25 parts of styrene but, as we shall see later, this ratio does not determine the ratio actually present in the final copolymer which is a function of reaction variables as well as the initial ratio of the ingredients. Consequently it was necessary to examine the composition of the final copolymer and to control it at the proper ratio of butadiene to styrene. The chemical composition was not the only factor to be controlled, however, since as we have seen, the properties of polymers unlike ordinary chemical compounds depend as much if not more on the chain structure. This is of course not only dependent on the nature of the starting ingredients but also on the manner in which they are combined into the chain.

At the time intensive work was undertaken in this country on Buna S type synthetic rubber little attention had been given to its characterization by physico-chemical means. The usual physical testing procedures involving the preparation of compounds by mixing in pigments and vulcanizing were of course being employed to supply useful information about the

copolymer produced and the vulcanization properties possessed by it. What was needed, however, were more precise and revealing tests, and tests which could be carried out directly on the copolymer itself. No ordinary chemical methods such as are applicable to the usual type of synthetic chemicals apply, for reasons which should be evident from our previous discussion. New methods of characterization designed to insure uniformity and satisfactory quality in the GR-S copolymer were required.

The precise and early control of the copolymer was of utmost importance. Non-uniformity in the product may cause serious troubles in fabricating operations such as are employed in tire plants, wire coating factories, adhesives manufacture, etc. Furthermore, with a varying product it often cannot be determined whether the trouble, when it occurs, is in the copolymer or in the method of fabrication being used.

What are the characteristics which must be controlled to insure a satisfactory product? To answer this question it was necessary to investigate a variety of GR-S copolymers and to conduct service tests on them in order to determine their practical performance. Some of these tests, particularly those on tires, have been very extensive. Some of the characteristics of the copolymer which experience has taught should be measured and controlled are:

1. The over-all or average styrene content in the butadiene-styrene copolymer.

   This necessitates (1) a method of separating the pure copolymer (which is the rubber-like component) from non-rubber components such as soap, salts, insoluble matter etc., and (2) a suitable method for determining the styrene content of the purified copolymer.

2. The percentage soap, fatty acids and low molecular butadiene-styrene compounds in the rubber.

3. The amount of "gel" fraction, if present, and the swelling volume of the gel.[7]

4. The average molecular size of the "sol" or soluble fraction of the copolymer.

5. The degree of branching of the sol molecules.

6. The molecular weight distribution of the sol.

In addition to the above tests on the final copolymer, control tests which can be used during the polymerization to tell when the reaction has progressed to the proper point were needed. In the following paragraphs we will take up in some detail the problem of characterization and attempt to show the basis on which methods have been evolved to control some of the quantities listed above.

Composition of GR-S and Its Determination

Given a piece of GR-S synthetic rubber, our first task from the standpoint of determining its chemical composition is to separate the pure copolymer which is responsible for the rubber-like properties from the non-rubber constituents. The latter comprise soaps or other emulsifying agents, fatty acids, salts, antioxidant and low molecular weight, non-rubbery butadiene-styrene products to the extent of several percent. Some of these minor ingredients, like the antioxidant, are essential whereas others play no important role subsequent to polymerization. All, however, must be separated from the copolymer before it can be properly evaluated. The analysis for the non-rubber components after separation is fairly straightforward and standard and will not be gone into here.

It has been found that the azeotrope of toluene and ethyl alcohol which consists of approximately 30 parts by volume of toluene to 70 parts by volume of alcohol is an excellent extractant for the non-rubber compounds and hence may be used to effect a separation[8, 9]. The procedure for isolating the copolymer is simply to place a quantity, say 10 grams, of the GR-S in an extraction thimble supported in an extraction flask as shown in Fig. 2. Another, more rapid, procedure is to reflux the azeotrope over the rubber for two hours, when extraction has been found to be essentially complete. This method is now used in the Standard Specification for all GR-S. The pure copolymer, left as residue, is the product to which we now turn our attention.

As has been mentioned, the ratio in which butadiene and styrene are employed in the starting mixture does not determine either the ratio in the whole copolymer at a given stage of reaction or the ratio present in any given chain molecule of the copolymer. Therefore the starting ratio cannot be relied upon to control the composition of the final copolymer. Experiments show that under certain process conditions large differences in composition between different fractions of the copolymer do occur. Even under the best conditions theoretical considerations predict that variations must occur between molecules since the ratio of the reactants is continuously changing during the reaction.

Let us examine the chemistry of the process for a moment to try better to understand why these variations are possible. When styrene (S) reacts with itself polystyrene (S_x) is formed. Analogously polybutadiene (B_y) is formed in the case of butadiene (B). In GR-S both styrene and butadiene react to give a copolymer.

When a quantity of styrene undergoes polymerization, a distribution consisting of various numbers of long chain molecules of various lengths is formed. Thus, if we start with N_1 molecules of styrene, S, the polymeriza-

tion reaction results in the formation of molecules of polystyrene by the addition of S to S in chain fashion. The result may be expressed as follows:

$$N_1S \rightarrow a_1S_1 + a_2S_2 + a_3S_3 + \cdots + a_nS_n \qquad (4)$$
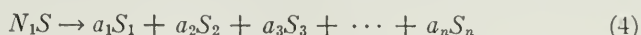
where each term represents a group of styrene molecules containing $1, 2 \cdots n$ styrene units, n assuming values up to several thousand depending on the

Fig. 2.—Apparatus for extracting non-rubber components from GR-S.

reaction conditions. If N_1 is very large there are of course many molecules, a_n, formed of the length corresponding to each value of n. In fact, $a_1 + 2\,a_2 + 3\,a_3 + \cdots + na_n = N_1$. The first term in (4) allows for the molecules which do not react, the second represents the dimers, the third the trimers, etc.

In an analogous way we may consider N_2 molecules of butadiene, B, to polymerize into chain molecules of various lengths:

$$N_2B \rightarrow b_1B_1 + b_2B_2 + b_3B_3 + \cdots + b_nB_m .\tag{5}$$

Now if styrene and butadiene molecules react together, as in the production of GR-S, we can represent their copolymerization as the insertion of the styrene chains (or portions of them) of (4) at random points in the butadiene chains of (5) to form chains S_jB_k. That is

$$N_1S + N_2B \rightarrow \Sigma \ a_iS_jB_k\tag{6}$$
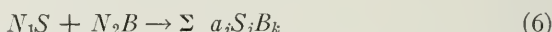
where j and k take on a variety of integral values and in any particular chain the arrangement of S and B units is probably random.

In practice, in the reaction represented by (6), N_2/N_1 has the value of approximately 6 since 75 parts by weight of butadiene are employed to 25 parts of styrene. Each chain molecule therefore would be expected to contain about 6 butadiene residues to each one of styrene. It is actually found, however, as indicated above that the starting ratio is not adhered to throughout the reaction, the molecules formed early being richer in butadiene and those formed later being poorer in butadiene than the starting ratio of 6 to 1. But, not only is the ratio B_k/S_j a variable from molecule to molecule of the copolymer formed but also their sequence along the chain is variable. Thus, in equation (6), even when equal numbers of styrene and butadiene molecules are present, a strict alternation is apparently not maintained but "strings" of one pure component or the other, form.

In the GR-S reaction the *weight* ratio of butadiene to styrene in the first molecules formed may be as high as 4:1 or more from a starting charge of ratio 3:1. Thus, the average weight percentage of styrene in the GR-S copolymer first formed is about 8% below that in the original charge (25%) and increases with conversion so that at the point where the reaction is stopped the copolymer forming contains about 29% styrene. Analogously there is evidence to show that in GR-S no regular sequence of butadiene and styrene along the chain molecules exists but rather a more or less random entrance of the two residues into the molecules with a frequency approximating the 6 to 1 ratio, as the extent of combination (percentage conversion) of the two ingredients approaches completion where obviously the two must become equal. Figure 3 illustrates this behavior for a typical sample prepared in the laboratory. An integral curve showing the cumulative percentage styrene and a differential curve representing the percentage styrene in the increment of the copolymer are illustrated.

It must be left to future research to determine how important the molecule to molecule variations in styrene content are in terms of useful properties

and to devise ways of eliminating them if necessary. For the present, we are perhaps justified in assuming that these variations can be neglected. The control considered here therefore relates to the over-all or average composition of the copolymer.
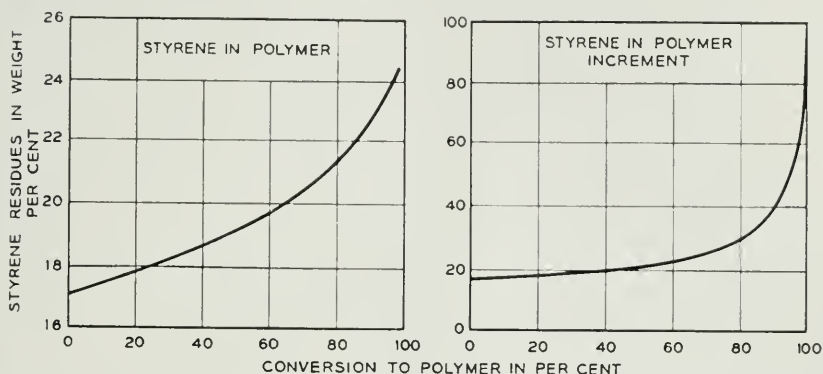
Fig. 3.—*Left:* Cumulative percentage by weight of styrene in the copolymer as a function of percentage conversion for an initial 25 percent styrene charge. *Right:* Percentage by weight of styrene in the polymer forming at any instant as a function of conversion for an initial 25 percent styrene charge.

DETERMINATION OF STYRENE CONTENT

Many suggestions involving both chemical and physico-chemical methods for measuring the average styrene content of GR-S copolymers have been proposed. Physico-chemical methods when applicable have an advantage in speed and precision over straight chemical methods and therefore have been more carefully examined. Both ultra-violet absorption[10] and refraction[8] have been shown to be applicable but since the absorption method is much more sensitive to impurities, the refraction method has proven the most general. It has the advantage also that it can be employed with polymers containing considerable gel fraction.

The refraction method is based on the fact that the styrene residues in the copolymer provide a greater contribution to the refraction of light passing through the solid or a solution of the solid than do the butadiene residues. Early work at the Bell Laboratories showed that the determination of the refractive index of the solid *unpurified* copolymer led to errors. In addition, the determination of the refractive index even of purified polymers was not precise if much gel was present, as frequently was the case with the early synthetic product. As a consequence a method, based on the use of the interferometer, was developed[8, 10]. The procedure is to disperse 2.4 grams of the pure copolymer in benzene, transfer the contents to the interferometer

cell and make a reading of the change in refraction compared to the pure benzene. This value, with the help of a curve relating styrene content to refraction, enables the true styrene content to be determined. The curve of refraction as a function of styrene content must be constructed beforehand and is shown in Fig. 4. This curve is obtained by measuring the refraction of pure polystyrene on the one hand and polybutadiene on the other. Checks also were made by independent methods of estimating composition in the range of the usual Buna S-type synthetic rubber.
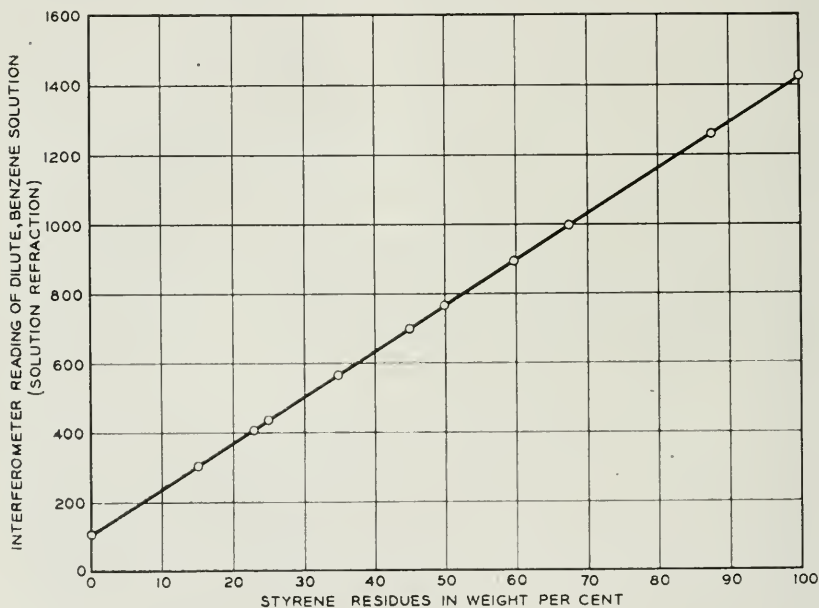
Fig. 4.—Refraction as a function of styrene content for solutions in benzene of polymers containing known percentages of styrene.

Through the use of this method it has been possible to control the styrene content of the copolymer to about ± 0.2 weight percent styrene residues, which is amply close for all purposes. Figure 5 shows the apparatus employed in this determination, the interferometer. More recently, it has been possible to employ a simpler procedure where a milling of the copolymer is introduced to remedy difficulties early encountered in the determination of the refractive index directly on the solid[11]. Although not as precise as the interferometer method, this method is shorter and as a consequence is finding application in process control. It is safe to say that today, with these methods, the control of the average composition of GR-S procuced in this country is now entirely adequate for all purposes.
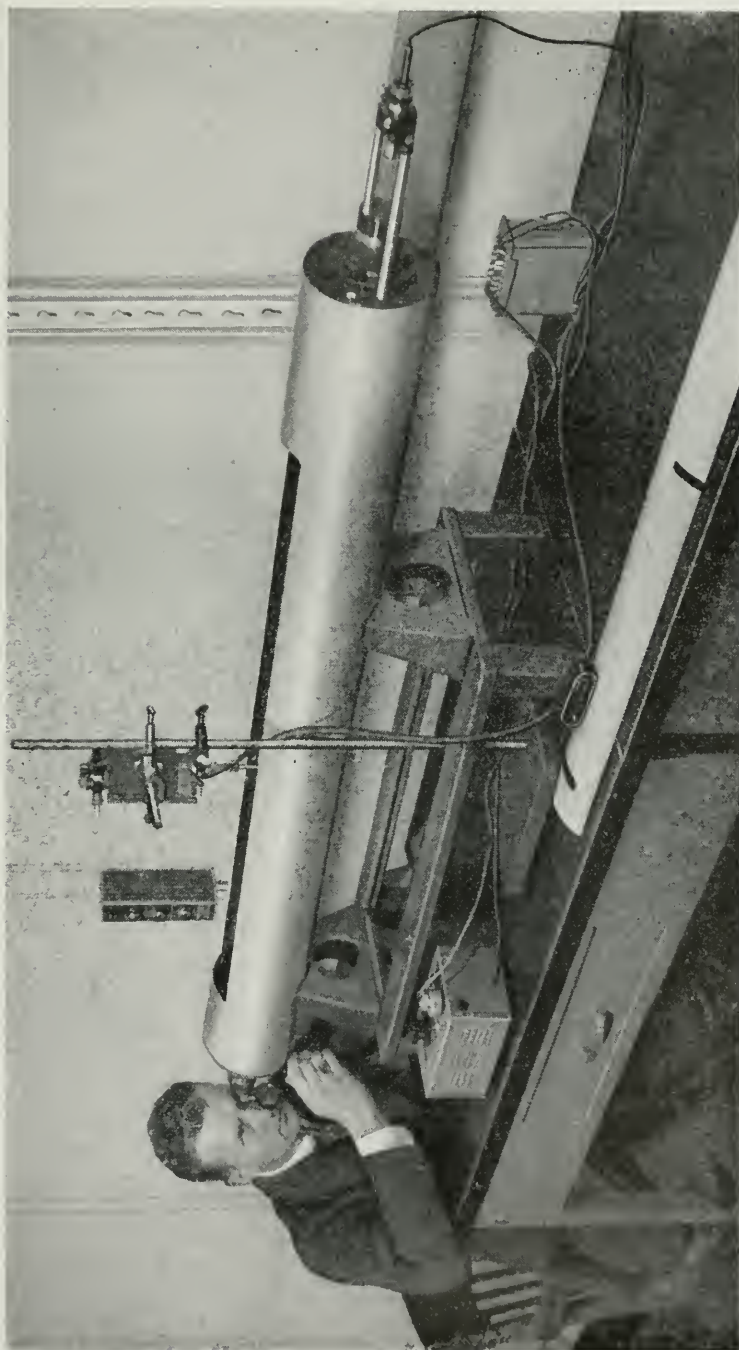
Fig. 5.—Interferometer used in the determination of styrene content of synthetic rubber from refraction.

Molecular Weight Distribution in GR-S

Unlike the linear polyesters whose molecular weight distributions can be calculated from simple assumptions (Fig. 1), the distribution of molecular sizes present in polymerization polymers cannot, at the present state of our knowledge at least, be accurately predicted. With linear polymers of uniform composition it is possible to determine experimentally the approximate molecular weight distribution by fractional precipitation of the dissolved polymer from dilute solution. This procedure, to yield good results, must be carried out under very careful control, and requires considerable time. The usual procedure is to prepare a solution of the polymer to be studied and add to it portions of a precipitant. The successive fractions of the whole polymer precipitated are then examined for average molecular weight by some suitable method. This procedure can give only a crude separation but often furnishes useful information. More accurate results require the use of very dilute solutions and the precipitation is best carried out by lowering the temperature to produce insolubility at each step. The experimental distribution curve is then obtained by plotting as ordinate the weight fraction and as abscissa the average molecular weight (weight average or number average) corresponding to each fraction. In this way an integral curve is obtained which on differentiation gives differential curves of the type shown in Fig. 1.

In GR-S, such a fractionation procedure is complicated by the fact that all of the molecules of the copolymer are not of the same type. For as we have seen we may encounter differences not only in structure between molecules but also in composition either of which alone will, independently of molecular size *per se*, influence solubility.

In fact experiments have shown that fractions separated from GR-S actually do exhibit differences in styrene content attesting to the special complications of determining molecular distributions in copolymers by this method. In spite of this, fractionations of GR-S have been made which no doubt have qualitative value. As a result of such experiments it has been found that molecular size distribution in GR-S is highly dependent on impurities present during the reaction as well as on other factors. When, however, the process and raw materials are suitably controlled it is likely that the shape of the curve does not vary greatly. Under these circumstances the number average molecular weight determined by osmotic pressure furnishes a measure of molecule size.

If the molecules are not too highly branched, we may employ viscosity measurements to furnish a "viscosity average" molecular weight. Since the latter measurements are the simplest to make they are generally employed[12], although care must be used to insure proper interpretation of results. In general, the average molecular weight given by the viscosity will

fall nearer to the low molecular weight end of the distribution curve than does the true weight average molecular weight. Only for a homogeneous system does it coincide with the number average value. Hence, the difference between the two can be used as a rough measure of the broadness of the distribution.

Light scattering from solutions offers possibly an absolute way of getting the true weight average value. If the molecules are small compared to the
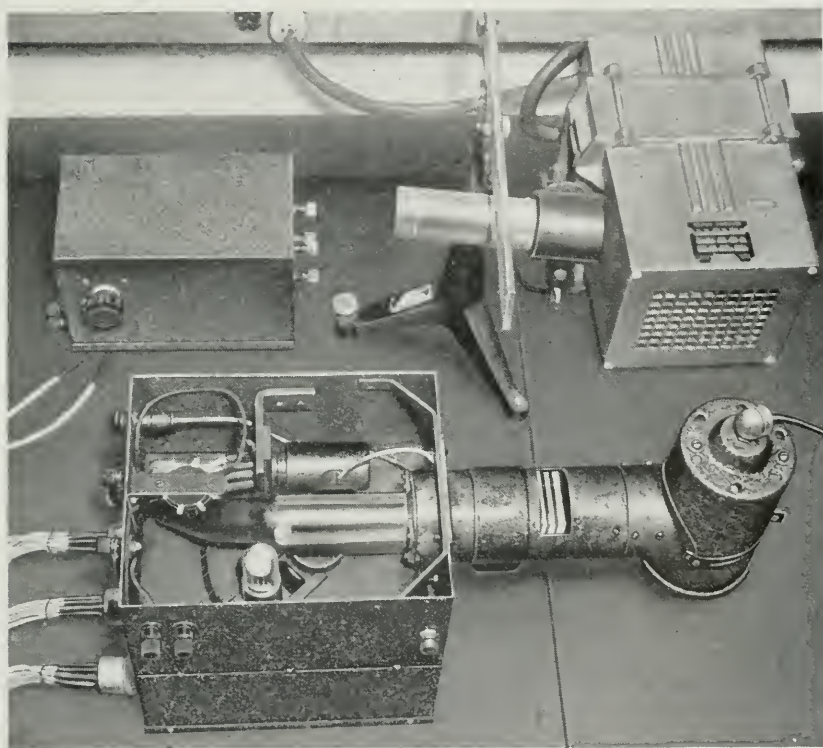
Fig. 6.—Apparatus for measuring the intensity of scattered light from solutions of polymers

wavelength of the light used and solutions of various dilutions are employed, measurements of turbidity τ, i.e. the fraction of the total light scattered per cm. of path, allow the weight average molecular weight $\overline{M}_w$ to be calculated according to

$$\overline{M}_w = \frac{1}{H(c/\tau)_o}$$

where H is a constant, c is the concentration and $(c/\tau)_o$ is the value found by extrapolation to zero concentration[13]. Further study of this method is

required before direct results can be obtained on GR-S. An apparatus employing electron multiplier tubes for measurement of the intensity of the scattered light, which was developed in the Laboratories, is being used to study this new technique. Its original form is illustrated in Fig. 6. Likewise, photographic determination of scattered light has established good correlation with independent molecular weight evaluation of certain other polymers[14].

The distribution of molecule sizes in GR-S has a profound influence on its properties. It is controversial still as to whether a uniform or non-uniform distribution is desirable for all considerations. The presence of low molecular material favors ease of processing but depreciates properties. High molecular material behaves the opposite. It is customary to regard the viscosity, either of the rubber itself or the dilute solution viscosity, as a measure of the average molecular weight. While this assumption is not wholly true, it is partly justified because the shape of the distribution curve as commonly measured for GR-S is roughly constant. When osmotic measurements can be made sufficiently accurately, the number average molecular weight together with the viscosity average provides a more precise measure of the distributions present.

Chain Structure of GR-S and Its Characterization

As for simple polymers, the chain structure of GR-S is best considered from two points of view: the micro-structure and the macro-structure. The microstructure, which has already been briefly discussed, is concerned with the kinds of atoms forming the chain, their arrangement in space and the manner in which they pack with the atoms of neighboring chain molecules. It is this structure which is all-important in determining the nature of the forces between molecules and, in turn, the intrinsic rubber-like properties of the polymer. The micro-structure also determines the chemical properties of the compound. The macro-structure, on the other hand, is not dependent on the kind of atoms in the polymer or their immediate relation to each other but with the length of the chain molecules, their general shape and the extent to which they are joined with the other molecules (netted) or what were originally other molecules in the material. It is the macro-structure which plays the chief role in plasticity and viscosity of the rubber during processing, its smoothness or roughness during extrusion, the extent to which it elongates or creeps on stretching and the extent to which it swells in solvents.

Let us approach the problem of the micro-chain structure of GR-S by considering the possibilities from the organic structural point of view. In the formation of GR-S about 6 butadiene molecules combine with each styrene molecule. In (9) butadiene is shown in brackets and styrene residues

are between them.   There are at least four important ways in which the chain structure of GR-S can deviate from the simplest structure, namely

$$
\cdots \left[\begin{array}{cc} H & H_2 \\ C & C \\ C & C \\ H_2 & H \end{array} \right]_6 \left[CH \begin{array}{cc} H_2 & \\ C & \\ \end{array} \right] \left[\begin{array}{cc} H & H_2 \\ C & C \\ C & C \\ H_2 & H \end{array} \right]_6 \left[CH \begin{array}{cc} H_2 & \\ C & \\ \end{array} \right] \left[\begin{array}{cc} H & H_2 \\ C & C \\ C & C \\ H_2 & H \end{array} \right]_6 \cdots \quad (9)
$$

which would arise from a regular addition of butadiene and styrene.   In the first place, as noted before, the styrene and butadiene units can be badly mixed up in the chain and not arranged in any special order.   Secondly, the butadiene or the styrene units may be reversed end for end in the chains. This will make no difference in the case of the butadiene provided the molecule has a center of symmetry but this is probably not the case.   Again, the butadiene residue may assume either the well-known *cis* or the *trans* configurations shown in (10) because of the double bond present in it.

$$
\begin{array}{cc} H & H_2 \\ C & C \\ \text{\textbackslash}C & C\text{\textbackslash} \\ H_2 & H \end{array} \qquad \begin{array}{cc} H & H \\ C{=}C \\ \text{\textbackslash}C & C\text{/} \\ H_2 & H_2 \end{array} \qquad (10)
$$

<center>trans unit   cis unit</center>

Finally, the butadiene unit may be combined into the chain as a 1,2 or as a 1,4 unit.   In the former case a vinyl group is appended to the chain molecule whereas in the latter it is absent:

$$
\begin{array}{c} H_2 \\ C \\ H \,\text{/} \\ C \\ | \\ CH \\ \| \\ CH_2 \end{array} \quad \text{or} \quad \begin{array}{cc} H & H_2 \\ C & C \\ \text{\textbackslash}C & C\text{\textbackslash} \\ H_2 & H_2 \end{array} \qquad (11)
$$

<center>1,2 unit   1,4 unit</center>

If these possibilities are considered, one representation of the chain molecule over a distance we should consider within the definition of micro is as follows:

$$
\begin{array}{c}
\text{H}_2 \quad \text{H} \quad \text{H}_2 \quad \text{H}_2 \quad \text{H} \ \text{H} \quad \text{H}_2 \quad \text{H}_2 \\
\text{C} \quad \text{C} \quad \text{C} \quad \text{C} \quad \text{C}=\!=\!\text{C} \quad \text{C} \quad \text{C}
\end{array}
\tag{12}
$$

Obviously many other combinations are possible which are even more involved.

There is considerable chemical as well as physical evidence to support the presence of all of these possibilities in the GR-S molecule. It is probable that the butadiene and styrene units enter the chain in an irregular manner, although, as we have seen, one molecule may acquire more total styrene or butadiene than another. The occurrence of *cis* and *trans* forms and head-to-tail arrangements is also irregular. The 1,2 and 1,4 butadiene structures likewise may occur randomly although the amount of 1,2 structure appears to vary somewhat depending on the type of reaction. It is not possible to review here the detailed evidence for the randomness and for the occurrence of these various features. The fact that x-rays when diffracted from stretched or cooled samples of GR-S fail to show evidence of crystalline or even of imperfectly crystalline material is proof that a disordered chain structure exists. X-rays, however, do not specify the cause of this disorder.

Work on synthetic linear polymers of known composition has demonstrated that relatively minute structural changes are able to cause marked disorder in polymer systems[16, 17]. It is not surprising, therefore, to find that GR-S copolymer is disordered. The important question is: what effect has the disorder on the properties and, if it is deleterious, what can be done about improving the chain structure? Without going into detailed arguments there is good reason to believe that an ordered chain structure is desirable for the best properties in a rubber. Only then is it possible for the chain molecules to pack together into crystalline-like regions on stretching and thus provide the resistance to tearing and breaking that are required. Natural rubber possesses this characteristic to an outstanding degree and polychloroprene and polyisobutylene when vulcanized also show considerable crystalline behavior on stretching. Other factors, such as the rate at which crystalline regions develop, are likewise important[17]. But the crucial requirement for toughness is the development of the crystalline type of forces on stressing.

It must be admitted that no great progress in reducing the chain disorder of GR-S has been attained as yet. Obviously, complete order because of the hybrid nature of the polymer is impossible. This was realized at the

outset of the research program and for that reason emphasis was placed on improving the macro-structure where obvious changes could be effected. We shall consider this phase of the work next.

We have already seen how the chain molecules of GR-S vary in size and in composition. They may vary also in over-all shape. Branching and cross-linking leading eventually to net-work formation may result during the chain growth or termination reactions. In this way variously shaped molecules may arise. Obviously the situation may become very complex and in reality we may have to do with mixtures where all types of molecular species are present at once.

What influence on the properties of the final compounded and vulcanized rubber do these various branched and netted chain structures have? It was not recognified at first that the gel part of GR-S was particularly different from the sol in its effect on ultimate properties. This was because no reliable measurements of sol or gel had been made and because sol and gel behaved differently during the compounding and processing steps[7, 12]. Some workers also did not appreciate that natural rubber and GR-S behave very differently in regard to the effect of processing on their ultimate properties.

It has since been established that the sol-gel properties are of importance both in the processing and in the final properties of GR-S synthetic rubber. It turns out that the amount of the sol and its molecular weight distribution and the amount of the gel and its swelling volume, which is a measure of the intensity of netting, enables us to make predictions as to what properties a given sample of rubber will exhibit during processing and in the final product[18]. This does not mean that other features of the sol and gel are unimportant. For example, methods of estimating the degree of branching (by means of concentrated solution viscosity)[6, 12] of the soluble portion have been worked out which undoubtedly will be useful if a more refined control proves desirable.

It is possible to make GR-S type rubber which is completely soluble. Such a product requires to be characterized only as to molecular weight distribution, composition and perhaps degree of branching. If the distribution of sol is such that there is an excess of low molecular material, the copolymer besides being soft and difficult to handle, provides cured stocks which have low tensile strength, poor tear and abrasion resistance, poor resistance to the growth of cracks and high hysteresis loss. If, on the other hand, an excess of high molecular material is present in the sol the copolymer is very stiff[18] and cannot be handled in the subsequent compounding and processing procedures. Aside from this difficulty its ultimate properties seem to be superior the higher the average molecular weight. When all considerations of properties and processing requirements are taken into

account a copolymer containing as nearly linear molecules as possible and having neither an excess of high or low molecular fraction is probably preferable.

Gel GR-S, depending on its swelling volume (see below), is a tough material totally lacking in plasticity. Swelling volumes as low as 10 are hardly distinguishable from vulcanized gum GR-S and in fact resemble it structurally because vulcanization is actually a special kind of gel formation. Ordinarily the swelling volumes of gel in GR-S range between 20–150[17, 18].

Commercial GR-S may contain both sol and gel, although the trend is to eliminate gel altogether. When large amounts of gel of moderate swelling volume are present the product is hard to mix, although it may extrude smoothly, and after processing, particularly if done hot, it is likely to give products which have higher modulus than copolymer free from gel, and to show poor resistance to cutting and crack growth—properties of great significance in tires and other applications.[18] It is therefore important that we should be able to determine sol and gel in the presence of each other. This need is particularly great in the case of characterization of copolymers after they have been subjected to processing and compounding[18]—treatments which often are responsible for profound changes in its molecular structure.

Methods of Characterization of Sol and Gel

Considerable work has been done at the Laboratories on methods for determining the sol-gel properties of polymers and in investigating the effects of various after treatments of the copolymers on their sol-gel characteristics[6, 18]. Figure 7 shows the type of apparatus employed for effecting the sol-gel separation[7]. The weighed copolymer sample is thoroughly dried, cut into small pieces and distributed on stainless steel screens contained in the bulb of the apparatus. About the 100 ml. of benzene is added and the parts assembled. After 24 hours or more standing without disturbance, the benzene containing the soluble part of the copolymer is carefully withdrawn by opening the stop-cock very slightly. The weight of the swollen gel left on the screens is obtained from the difference between the weight of the assembly after draining off the solution and its original weight. This divided by the original weight of the unswollen gel gives the swelling volume (SV) of the material. The slight density correction can be neglected.

The dilute solution viscosity is determined directly on 5 cc. of the solution withdrawn from the vessel and is calculated from equation (2). The concentration c is determined by evaporating a known volume of the solution and weighing the solid left after evaporation of the benzene. Figure 8 shows the viscometers and bath employed for the measurements of the relative viscosity. A variation of the dilute viscosity method adopted for
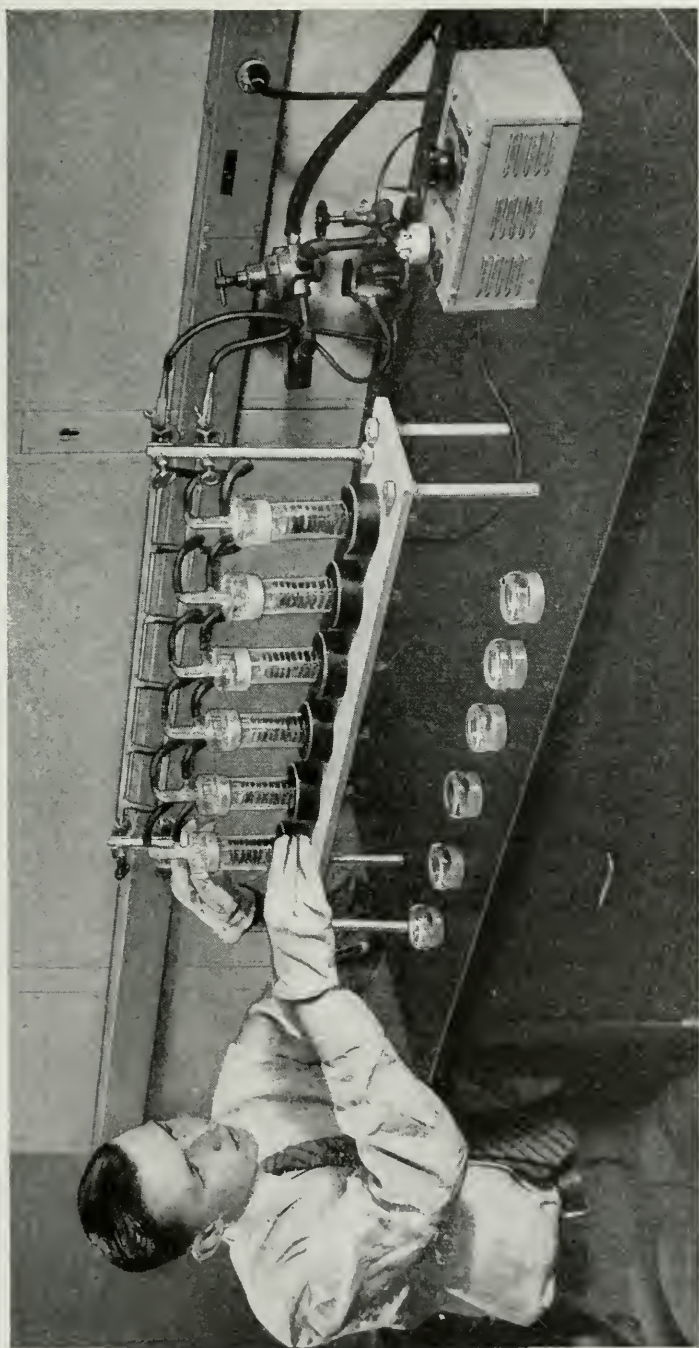
Fig. 7.—Apparatus employed in the determination of the sol-gel content of synthetic rubber.

use directly on latex has been employed as a control during the synthesis of GR-S[19]. This test, referred to as the vistex test, consists in adding 1 ml. of the latex sample to be examined to 100 ml. of a solvent having both hydrophobic and hydrophilic properties, such as a mixture of 70 parts (by volume) of xylene with 30 of pyridine, or 60 of benzene and 40 of *t*-butanol. The clear solution is run through the viscometer in the usual manner and the relative viscosity used as a measure of extent of reaction. The test has the
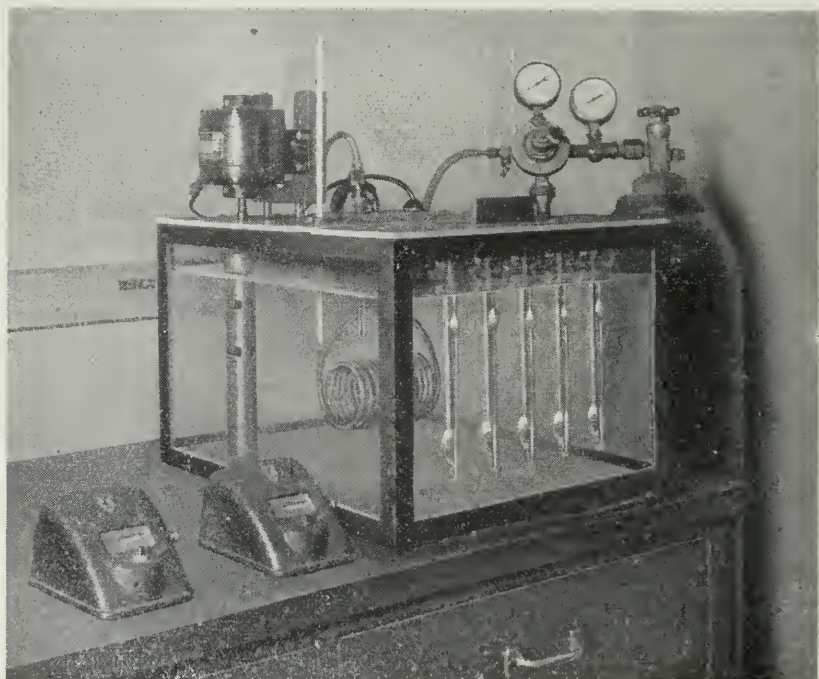
Fig. 8.—Viscometers and bath used for the determination of dilute solution viscosity of polymer solutions.

advantage of great speed, thus providing control of the reaction, step by step. Figure 9 shows the apparatus employed in the determination of concentrated solution viscosity (*CSV*).[20] In this measurement a 15 percent solution of the copolymer in xylene is made by weighing the required quantity of GR-S into a test-tube adding the precise volume of xylene and stoppering. The solution is homogenized by moving a steel armature through it in the test-tube by means of a strong electro-magnet. A trace of acetic acid is added to eliminate thixotropic effects. After complete dispersion has been effected the viscosity is determined by the falling ball

method. Branched copolymers show inordinately high concentrated solution viscosities. The latter may therefore be employed as a measure of degree of branching or approach to gelation when supplemented by dilute solution viscosity measurements. Furthermore, the power required to maintain the armature stationary as measured by the current passing through the magnet furnishes data useful in predicting how a given copolymer sample will process.
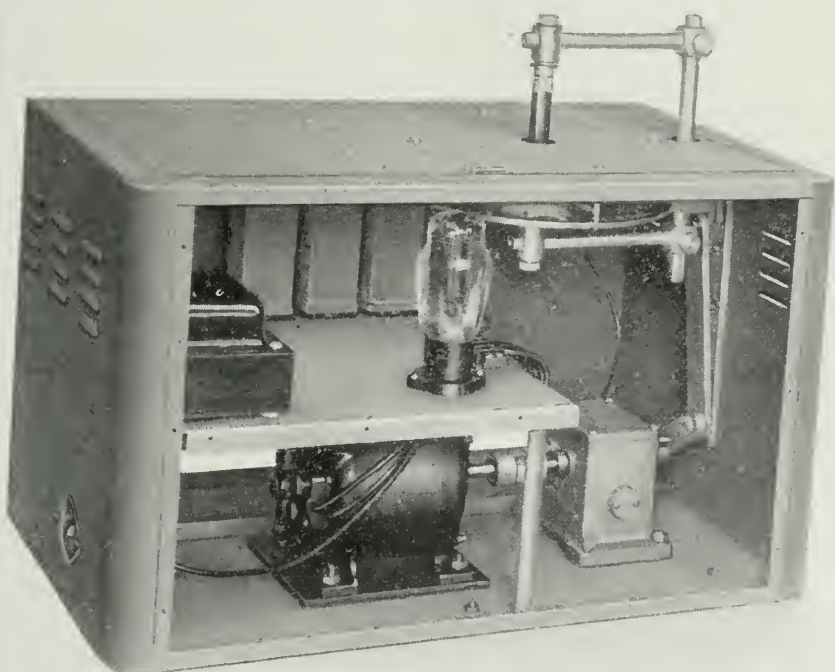
Fig. 9.—Apparatus employed to effect the solution of synthetic rubber prior to the determination of concentrated solution viscosity.

APPLICATION OF SOL-GEL METHODS TO CONTROL PROCESSING

In addition to their application to the control of synthetic rubber in production, the sol-gel methods of characterizing the copolymer which have been briefly described above are of very great use in elucidating what happens during the processing of the rubber[18]. By the term "processing" is meant the operations which are carried out on the copolymer subsequent to its manufacture and prior to its vulcanization into its final form. These operations involve working the rubber on machinery (plastication) in order to render it soft and satisfactory for mixing in pigments and for extrusion
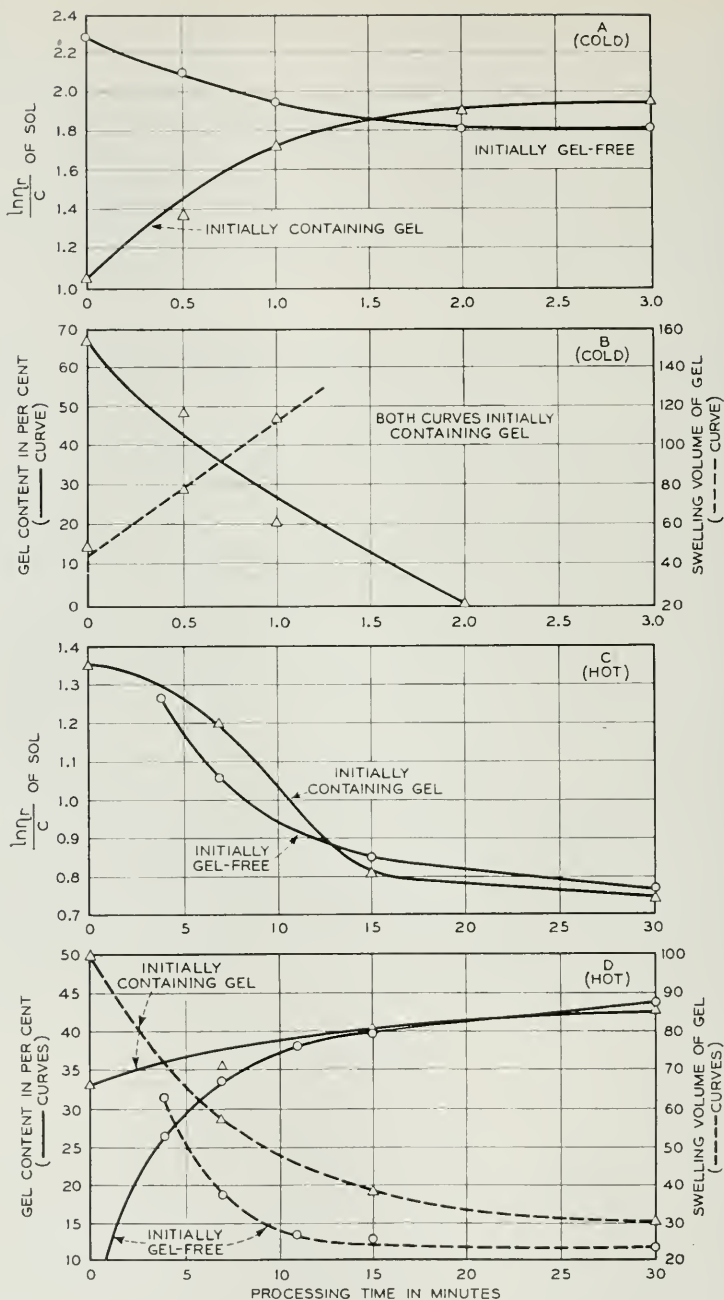
Fig. 10.—Curves showing change in solution viscosity $\left(\dfrac{\ln \eta_r}{c}\right)$ of sol, gel content, and swelling volume of gel with time of milling. Both cold (A and B) and hot (C and D) milling are shown for samples of synthetic rubber containing no gel and gel of low swelling volume.

378

and molding.   Several different types of machines are employed including mills, calenders, Banbury mixers and extrusion machines.   At every stage and particularly in Banbury mixing, where carbon black is generally mixed in, the copolymer undergoes changes which affect its performance in the finished product.   This is especially true if, as nearly always happens in practice, considerable heat is developed during the operation.   Indeed, it is frequently true that processing operations have more to do with the ultimate rubber properties than do factors in the production of the copolymer itself[18].   It is only fairly recently that this point has been sufficiently emphasized and considerable progress made in controlling the processing steps to the same extent as the polymerization is now controlled.

To see what happens during processing of the copolymer let us assume we subject two extreme types of GR-S, one containing no gel and one containing gel of rather low swelling volume, to a typical hot processing treatment consisting of hot mastication and Banbury mixing in of carbon black.[18]   In addition, in order to exhibit differences in processing let us consider the effects of cold processing on the same two samples.   As is evident from Fig. 10 which summarizes the results,[18] hot processing tends to build up gel and decrease its swelling volume in both of the rubber samples.   The dilute solution viscosity on the other hand falls.   This behavior although of advantage to subsequent extrusion and calendering operations is definitely opposed to securing the best mechanical properties in the final rubber. Cold processing has the opposite effect on the samples.   Thus, the copolymer not containing gel is little affected, whereas the gel in the other is broken down and gives rise to a higher dilute solution viscosity.

In processing, therefore, important changes in the chain structure of the copolymer are brought about.   Under certain circumstances, these are beneficial, but since most processing involves considerable heat development the changes are usually detrimental.   It is of the utmost importance therefore that a type of copolymer is produced which is compatible with the type of processing machinery already installed in industry.   In addition, uniformity of the copolymer is of very great significance if control of processing operations is to be achieved, for such control cannot be attained with a variable starting material.

The processing step which involves the mixing in of carbon black (or other pigment) is perhaps the most important.   Unfortunately, the presence of the carbon black makes it impossible to employ the usual sol-gel analysis because a new kind of insolubility enters.[18, 21]   In addition to the primary valence gel discussed up to this point, a secondary valence combination involving the carbon black and the large sol molecules forms.   This is not immediately distinguishable from the first type of gel unless we have other reason to know that the latter is absent and does not form during the

mixing operation. This phenomenon of the insolubilization of natural rubber by carbon black has been known for some time.[22] Only recently, however, has its relation to the structural features of GR-S copolymer become apparent.[21] Work is now underway to allow an estimation of both types of gel in the presence of one another. When this is achieved the analysis of the reactions occurring during compounding will be further facilitated.

CHAIN STRUCTURE AND POLYMER PROPERTIES

We have now reviewed some of the molecular complexities which are involved in the synthesis of GR-S synthetic rubber. It remains to discuss more in detail the influence of chain structure on the properties we associate with rubber-like behavior. We might begin by asking ourselves two questions: (1) What makes a polymer exhibit rubber-like properties? (2) What composition and chain structure are desirable in a rubber? The first question involves a discussion of the theory of rubber-like elasticity. The answer to the second involves an inquiry into the specific use to which the material is to be applied. Since most of our rubber is employed in tires let us consider the special requirements for that use.

Taking up the first question, we fall immediately into the pit of having to define what a rubber is and how it differs from a plastic. Originally rubber meant "natural rubber". When synthetics with rubber-like properties appeared we adopted the term "synthetic rubber" to describe them. Some have objected (unsuccessfully) to the use of this term because it implies synthetic natural rubber and have proposed the word "elastomer" instead. Others have gone still further and suggested other terms (usually ending in mer) for various plastics and rubber-like materials.

All of these new names seem unnecessary. Polymer is the inclusive term. The term rubber simply has come to mean a polymer which at ordinary temperatures has properties like natural rubber. A plastic is a polymer which at ordinary temperatures is hard and which usually becomes soft and deformable at higher temperatures. Such terms as rubber-like plastic or glass-like plastic are frequently employed. This kind of terminology is admittedly loose but it often tells just as much in familiar words as does the newly proposed nomenclature.

The significant fact is that there is a perfectly consistent and orderly relationship between the properties of polymers and their chemical composition and structure. Fundamentally, the major factor which determines whether a long chain polymer will be a rubber or a plastic is the magnitude of the forces acting between chain molecules. If the forces between polymer molecules are low, the polymer is a rubber; if they are high, it is a plastic. And obviously since these forces can be regulated nicely there are all grada-

tions from the hardest to softest polymer. A rubber therefore may be regarded simply as a soft plastic—one in which the forces between chains are very low—with one important distinction, namely that soft plastics to show rubber-like properties must be "vulcanized" i.e. a few very strong inter-chain linkages must be established to prevent slippage. Some plastics can be vulcanized, too, but here the inter-chain forces are high anyway and the few additional strong bonds are not essential. However, the usual inter-chain forces in plastics being of the van der Waals' type are very susceptible to temperature. Consequently, if we weaken them by raising the temperature we can, provided the plastic is "vulcanized", cause it to acquire rubber-like properties at the higher temperature. So the distinction between rubbers and plastics is in the last analysis slight.

We still have not answered the question as to what causes a polymer to exhibit rubber-like properties. In fact it was only during the last 10 years or so[23] that the answer has been known, which is surprising, because it is simply "temperature". Contrary to previous views, forces between atoms in the same chain have little to do with the long range retraction phenomenon shown by rubbers, at least at elongations up to about 200%. The stretched polymer returns because of the thermal heat motion in the mass which seeks to restore the elongated chain molecules to their more stable, kinked-up configurations. The molecules, through their vulcanization points, communicate their retraction behavior to the entire mass. Thus theory agrees with experience that the chemical constitution of the polymer is of secondary significance. As long as chain molecules are present which are capable of kinking-up by rotations about chemical bonds, as long as the forces between molecules are not large compared to the thermal energy, and as long as the molecules are interconnected at points so as not to slip, we shall have a rubber-like substance whether we call it a rubber or a plastic or an elastomer.

Coming to our second question, it might now take the form: what composition and chain structure are desirable in a rubber for tires? We shall see that the qualitative views expressed above must be altered if we are to explain the more intimate properties of rubber involved in this application. We have already seen from the sol-gel discussion what some of these refinements are. There are certain differences between GR-S type synthetic rubber and natural rubber, however, which go back even farther and involve the manner in which the molecules pack and slip over one another during deformation. These have geen discussed above under micro-structure.

Man learns largely by imitation and our knowledge of rubber-like behavior has been no exception. Natural rubber possesses amazing qualities which no synthetic product has yet been able entirely to duplicate, although we have found in many cases ways to overcome weaknesses in synthetic rubbers by round-about means. For example, the hysteresis loss in vulcanized

natural rubber, i.e. the heat generated per cycle of reversible stretching, is less than any synthetic rubber of comparable gum tensile strength. The resistance to crack growth of vulcanized natural rubber, particularly in gum, i.e. unfilled form, is better than any synthetic of comparable low-temperature flexibility.

What are the causes of these differences? We can make synthetic rubbers from hydrocarbons (in fact from isoprene itself) which judged simply from the composition should have the same interchain forces acting as in natural rubber. That this is not so is shown by the fact that polyisoprene in all the properties which would fit it for a tire rubber is much inferior to natural rubber. We are forced to conclude that it is the form of the individual isoprene units (*cis* or *trans*) and the way in which they are placed in the chain that determines.

In stretched natural rubber we are convinced that units are orderly arranged in the *cis* configuration as follows:

$$
\begin{array}{ccccccccc}
CH_3 & & CH_3 & & CH_3 & & CH_3 & \\
\diagdown \; H & & \diagdown \; H & & \diagdown \; H & & \diagdown \; H \\
C{=}C & & C{=}C & & C{=}C & & C{=}C \\
\diagup \quad \diagdown & & \diagup \quad \diagdown & & \diagup \quad \diagdown & & \diagup \quad \diagdown \\
-C \quad C{-}C & & C{-}C & & C{-}C & & C{-} \\
H_2 \quad H_2 \; H_2 & & H_2 \; H_2 & & H_2 \; H_2 & & H_2
\end{array}
$$
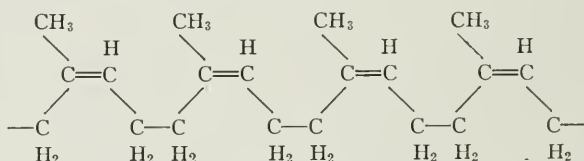
In polyisoprene on the other hand they are probably in random disorder:

$$
\begin{array}{cccccc}
CH_3 & & CH_3 & H_2 \; H_2 & & CH_3 \\
\diagdown & & \diagdown & C{-}C \; H & & \diagdown \; H \\
C{=}C & & C{=}C & C{=}C & & C{=}C \\
\diagup \quad \diagdown & & \diagup \quad H & \diagdown \quad \diagdown & & \diagup \quad \diagdown \\
-C \quad C{-}C & & C{-}C & CH_3 \; C{-}C & & C{-} \\
H_2 \quad H_2 \; H_2 & & & H_2 \; H_2 & & H_2
\end{array}
$$
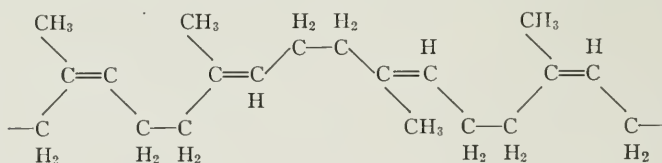
We might expect then that the interchain forces, which change so critically with distance, are less of a match for the thermal energy in synthetic polyisoprene than in natural rubber. The high degree of molecular uniformity in natural rubber, as X-rays show, gives rise to a crystallization on stretching which is entirely absent from polyisoprene (see Fig. 11). Even before crystallization as such has progressed far (it starts in nuclei and these multiply throughout the mass) the interchain forces have prepared for and compensated in an effective way for the increased effect of stress tearing the chains apart and the increased thermal energy tending to weaken the interchain forces. We may look upon natural rubber as a substance which progressively and automatically transforms itself to a plastic as it is elongated. It is these crystallization phenomena which are responsible for high gum tensile strength[17] and outstanding resistance to the growth of cracks.

To return to our question and ask once more what structure we desire in a rubber for tires we see that although we cannot quite write an order in terms of a chemical formula we can state general requirements. We want a polymer in which the interchain forces are as low as possible to give us low hysteresis. At the same time we want regularity of chain molecules to provide a minimum loss of cohesion with rising temperature and rising elongation. To a chemist this sounds like an order for natural rubber and the design of Buna S which we were forced to imitate in the emergency seems wrong. If research can iron out some of this irregularity, a further improvement in our product perhaps can be achieved. The chemist by the clever trick of adding styrene to butadiene has provided himself a way he can
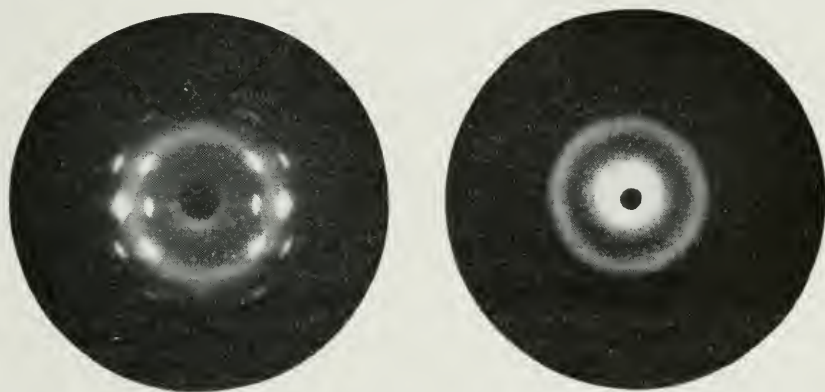
Fig. 11.—X-Ray photographs of natural rubber, stretched (left) and synthetic polyisoprene, stretched (right).

regulate the interchain forces and therefore the degree of rubberiness of Buna S. He is able to make it harder and stronger at will by increasing the amount of styrene, something nature is unable to do. But his task is not finished until he can control also the order and the packing of his molecules or devise some equally clever way of getting the interchain forces to behave.

Conclusion

We have attempted to review some of the problems arising out of the effort to achieve the best possible Buna S type rubber for our war emergency and to show how they have been attacked. We have also tried to give a simple account of some of the theories underlying the behavior of polymers. The story of synthetic rubber is of course much broader both in theory and practice than we have indicated.

Future synthetic polymers will be devised to meet the intimate requirements of many diverse applications. Engineering will be more precise and control of our materials will be based more on scientific methods. It is romantic to read from a recent popular book "you see him in his shirt sleeves cutting off a piece of rubber with his knife, smelling it, biting it and stretching it. Then he either looks satisfied or worried. Laboratory reports give him a complete report on the sample but a prodigious memory and a sixth sense born of years at his job often tell him whether the rubber will make a good tire." But this is hardly the way the future engineer will judge. It is hoped that the present account has helped to point out how scientific methods are being applied and how research can supply a safe guide to the wise control and application of synthetic materials.

Acknowledgement

References

1. Harries, "Untersuchungen über die Natürlichen und "Künstlichen Kautschukarten"," Berlin (1919).
2. Sponsler and Dore, *Colloid Symposium Monograph*, *4*, 174 (1926).
3. Staudinger, *Ber.*, *59*, 3019 (1926).
4. Vorländer, *Ann.*, *280*, 167 (1894)
5. Flory, *Jour. Am. Chem. Soc.*, *58*, 1877 (1936).
6. Baker and Mullen, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, March 23, 1943; *Ibid.*, June 24, 1943.
7. Baker and Mullen, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, Sept. 13, 1943.
8. Baker and Heiss, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, March 23, 1943; *Ibid.*, Sept. 14, 1944.
9. Baker and Heiss, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, Sept. 13, 1943.
10. Baker and Heiss, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, June 24, 1943.
11. Baker and Heiss, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, May 11, 1944.
12. Baker and Mullen, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, Nov. 23, 1943.
13. Debye, *Jour. Appl. Phys.*, *15*, 338 (1944).
14. Doty, Zimm and Mark, *Jour. Chem. Phys.* *12*, 144 (1944); *13*, 159, (1945).
15. Baker and Pape, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, March 14, 1944.
16. Baker and Fuller, *Jour. Am. Chem. Soc.*, *64*, 2399 (1942).
17. Baker, *Bell Lab. Record*, *23*, 97 (1945).
18. Baker and Mullen, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, June 12, 1944.
19. Baker, Mullen and Heiss, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, Jan. 21, 1944.
20. Baker and Mullen, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, Nov. 30, 1943; Baker, Walker and Pape, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, March 14, 1944.
21. Baker and Walker, Private communication, Bell Telephone Laboratories to Office of Rubber Reserve, Sept. 14, 1944.
22. Stamberger, *Kolloid. Z.*, *42*, 295 (1927).
23. Whitby, *Jour. Phys. Chem.*, *36*, 198 (1932).

Characteristics of Vacuum Tubes for Radar Intermediate Frequency Amplifiers

By G. T. FORD

1. Introduction

THE desired characteristics for vacuum tubes for use in broad-band intermediate frequency amplifiers are primarily high transconductance, low capacitances, high input resistance, and good noise figure. These characteristics determine the frequency bandwidth, amplification, and signal-to-noise ratio attainable with such an amplifier. The maximum operating frequency is generally limited by the input resistance of the tube which decreases as the frequency is increased, and, in some cases, by the tube noise which increases with increasing frequency. Three other characteristics which are also important are small physical size, low power consumption, and ruggedness. The present paper describes how these characteristics are related to the performance requirements for intermediate frequency (IF) amplifiers used in radar systems and shows how the requirements were met in the design of the Western Electric 6AK5 Vacuum Tube.

In a coaxial cable carrier telephone system of the type which was initially installed between Stevens Point, Wisconsin and Minneapolis, Minnesota[1] the upper frequency of the useful band is of the order of three megacycles per second (mc) and the bandwidth is of the same order. The Western Electric 386A tube was developed a number of years ago for amplifiers such as those used in this system. It is characterized by high transconductance and low capacitances. In radar receiving systems similar but more exacting requirements must be met for the IF amplifier. It is desirable to operate in many cases at a mid-band frequency of 60 mc, to have a bandwidth of the order of 2–10 mc, and to have as close to the ideal noise figure as possible. Additional considerations of great practical importance are low power consumption, small size, and ruggedness.

The choice of the mid-band frequency for the IF amplifier is influenced by considerations, a detailed discussion of which is outside the scope of this paper. For example, the characteristics of the beating oscillator and ite relation to the operation of the automatic frequency control (AFC) system, when AFC is used, are involved. The usual practice has been to

---

[1] "Stevens Point and Minneapolis Linked by Coaxial System," K. C. Black, *Bell Laboratories Record*, January 1942, pp. 127–132.
"Television Transmission Over Wire Lines," M. E. Strieby and J. F. Wentz, *Bell System Technical Journal*, January 1941, Vol. 20, p. 62.

standardize on a mid-band frequency of 30 mc or 60 mc. The latter frequency has been used in most radars developed at the Bell Telephone Laboratories.

In pulsed radar systems the transmitter is turned on and off by the modulator in such a way that radio frequency energy is generated for a pulse duration τ seconds at a repetition rate which is usually several hundred per second. When the transmitter is modulated by a pulse which approaches that shown in Fig. 1(a), the energy-frequency distribution in the transmitted signal is as shown in Fig. 1(b). Although the maximum of the distribution
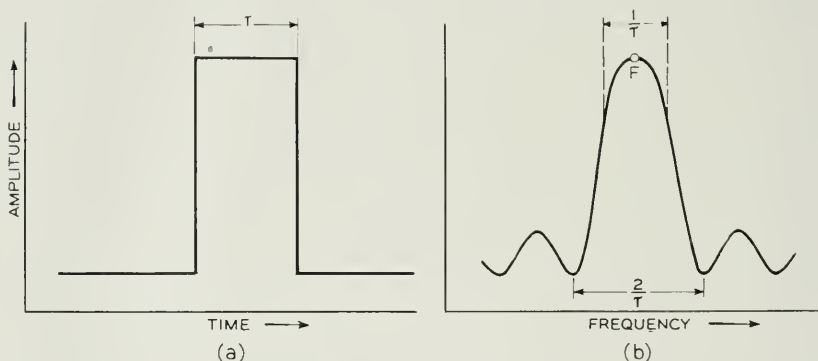
Fig. 1a—Modulating pulse
Fig. 1b—Amplitude—frequency distribution in transmitted RF pulse.

curve lies at the transmitter frequency F, there is considerable energy in the range $F \pm \dfrac{1}{2\tau}$ and the receiver usually has a bandwidth of at least $\dfrac{1}{\tau}$ in order to make as efficient use as possible of the energy in the echoes reflected by the target. For many radar applications, this requirement and the problems of transmitter and beating oscillator frequency stability result in the use of a bandwidth of as much as 10 mc for the IF amplifier.

The Western Electric 386A tube, Fig. 2(a), had characteristics which approached those needed to meet the IF amplifier requirements discussed above. It was therefore slightly modified in physical form for convenience of use and recoded the Western Electric 717A tube, Fig. 2(b). The 717A tube was used extensively in IF amplifiers in several radar systems. As the emphasis on small size and light weight for airborne radars increased, and, as the need for better characteristics became more pressing, further development was undertaken which resulted in the Western Electric 6AK5 tube, Fig. 3.

The importance of size and weight for radar systems to be used in airplanes is obvious. The use of miniature tubes in airborne equipment has
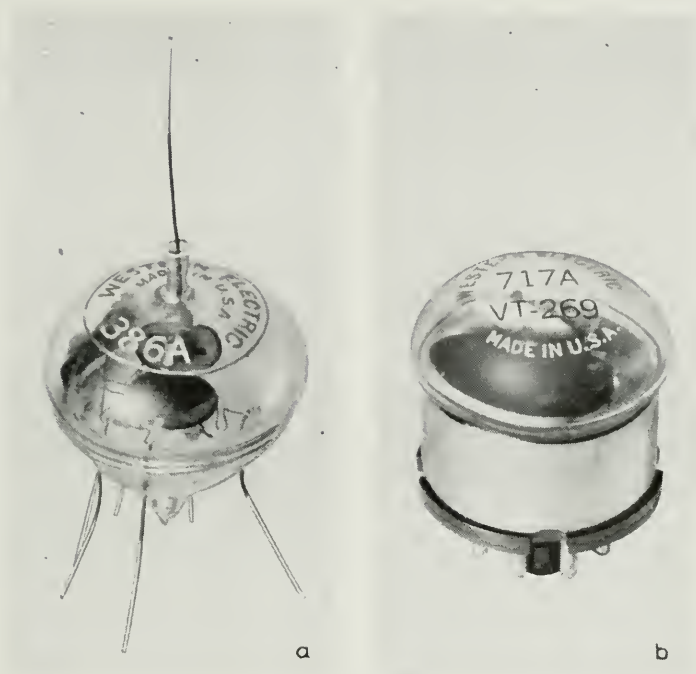
Fig. 2a—Western Electric 386A vacuum tube (full size).
Fig. 2b—Western Electric 717A vacuum tube (full size).

Fig. 3—Western Electric 6AK5 vacuum tube (full size).

been a consequence wherever their power handling capabilities are adequate to meet the performance requirements. The IF amplifier offers an ideal opportunity to effect substantial savings in both size and weight by using

small tubes.   When the IF frequency is as high as 30 mc or 60 mc, the
circuit elements are physically small, so that the tube size is relatively
important.   The power level is low enough so that miniature tubes are
applicable.   The photograph shown in Fig. 4 illustrates the reduction in
the size of the IF amplifier obtained by using 6AK5 tubes in place of 6AC7
tubes which were widely used previously.   The larger amplifier weighs 2
lbs. 4 oz. while the smaller one weighs only 9 oz.   Each of these amplifiers
provides an over-all amplification of about 95 db at a mid-band frequency
of 60 mc.   The bandwidths of the two amplifiers are comparable.   The
amplifier using 6AC7 tubes requires 31.3 watts of power, while the 6AK5
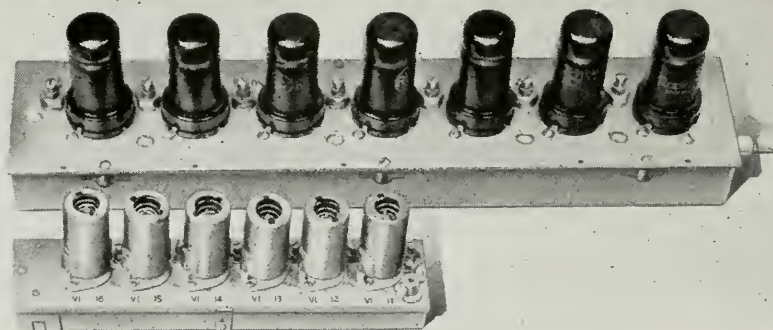
Fig. 4—60 Megacycle IF amplifiers (¼ full size).

amplifier requires 14.4 watts.   The power savings possible with the 6AK5
tubes are not particularly important from the power economy standpoint,
but rather because of the easier heat dissipation problem.   In compact
equipment such as airborne radar, the problem of keeping the operating
temperatures of the various components within safe limits is formidable.

In the later years of the war the 6AK5 tube became the standard IF
amplifier tube for radar systems and because of its superior properties was
used for other applications in many other radio equipments.

2. Amplification and Bandwidth

The amplification that can be obtained with a given number of stages,
and the useful bandwidth, are closely related.   Within certain limits, one
can be increased at the expense of the other.   In fact, the product of the
amplification per stage and the bandwidth is one important measure of the
goodness of a particular tube and circuit design.   A simple case will illustrate
how this comes about.   Assume the band-pass interstage shown in Fig. 5.
L is the inductance of the coil, R is the shunt resistance equivalent to the

load resistor and any loading effects due to high-frequency losses in the tubes or circuit elements, and C is the total shunt capacitance of the tubes, the circuit elements and the wiring. If it is assumed that R, L and C are

BLOCKING
CONDENSER
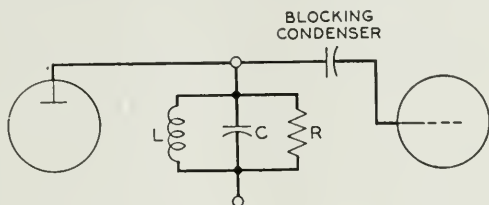
Fig. 5—Band-pass interstage.

independent of frequency, the magnitude of the impedance of this network is

$$|Z| = \frac{1}{\left[\left(\frac{1}{R}\right)^2 + \left(\omega C - \frac{1}{\omega L}\right)^2\right]^{1/2}} \dots\dots\dots\dots\dots\dots\dots\dots (1)$$

When $\omega = \omega_0 = \dfrac{1}{\sqrt{LC}}$ the impedance is a maximum and the frequency f_0

$= \dfrac{\omega_0}{2\pi}$ is the resonant frequency. The maximum value of the impedance is $|Z_0| = R$. If this type of interstage is used in a grounded-cathode type of circuit employing a pentode tube, the small-signal voltage amplification A_v from the control grid of the tube to the plate is

$$A_v = G_m |Z| \dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

where G_m is the grid-plate transconductance of the tube. It is assumed that $|Z|$ is small compared to the plate resistance of the pentode and that there is no feedback. The voltage amplification A_{v0} at the resonant frequency is then

$$A_{v0} = G_m R \dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

If the bandwidth is defined as $\Delta F = f_1 - f_2$, where f_1 and f_2 are the two frequencies where $A_v = \dfrac{A_{v0}\sqrt{2}}{2}$, it can be shown that

$$\Delta F = f_1 - f_2 = \frac{1}{2\pi RC} \dots\dots\dots\dots\dots\dots\dots (4)$$

The product of the mid-band voltage gain times the bandwidth can be called the band merit, B_0, and we have

$$B_0 = (\Delta F)(A_{v0}) = \frac{1}{2\pi RC}(G_m R) = \frac{G_m}{2\pi C} \dots\dots\dots\dots (5)$$

This expression for band merit has been derived from the product of the bandwidth and the voltage amplification for a simple band-pass interstage. Higher band merits can be realized with a given tube by using more complicated coupling networks.

Another interpretation of band merit is to say that it is the frequency at which the voltage amplification is unity. This is the frequency at which the product of the transconductance and the reactance of the shunt capacitance is unity.

From the foregoing expression for band merit it is evident that, in general, the higher the band merit the fewer is the number of stages that are required to obtain a given gain and bandwidth. It is highly desirable to keep the number of stages small in order to save space, weight and power consumption and to avoid the use of unnecessary components which reduce the

TABLE I

Type	Heater Power	Plate Current	Total Power Consumption	Nominal Transconductance	Transconductance per Unit Plate Current	Band Merit
	watts	ma	watts	umhos	umhos/ma	mc
6AC7	2.84	10	4.7	9,000	900	89.5
6AG7	4.10	30	9.6	11,000	367	85.3
6AG5	1.89	7.2	3.0	5,100	708	90.0
717A	1.10	7.5	2.3	4,000	533	71.4
6AK5	1.10	7.5	2.3	5,000	667	117.

reliability in operation. In practice the total amplification in the receiver is made high enough so that the system noise, with no signal, produces a fair indication on the output device when the gain control is set for maximum gain. For a bandwidth of 5 mc, the equivalent RF input noise power level is of the order of 2×10^{-13} watt and the power level necessary for a suitable oscilloscope presentation in a radar is about 20 milliwatts. The net over-all gain needed is then about 110 db. Making an allowance of, say, 15 db for losses in the detectors and elsewhere, a total of about 125 db gain is required. A minimum of about 110 db of this is usually in the IF part of the receiver. With this amount of gain as a requirement it is easy to see the importance of high band merit, small size, and low power consumption in the IF tubes.

Table I, above, compares the salient characteristics of the 6AK5 with those of other similar types of tubes.

The tube design factors which determine the band merit will now be examined by considering an idealized case. For an idealized plane parallel triode structure in which edge effects are assumed to be negligible, the plate current can be related to the tube geometry and the applied voltages approximately as follows:[2]

[2] "Fundamentals of Engineering Electronics," W. G. Dow, pp. 44, 102, et seq.

$$I_b = \frac{2.33 \times 10^{-6} A \left(E_{c1} + \dfrac{E_b}{\mu}\right)^{3/2}}{a^2 \left(1 + \dfrac{1}{\mu}\dfrac{a+b}{b}\right)^{3/2}} \quad \dots\dots\dots\dots(6)$$

It is assumed that the electrons leave the cathode with zero initial velocities. The failure to take into account the effects of initial velocities makes this equation only a fair approximation for close-spaced tubes, but it is still instructive to assume it is approximately correct for the purposes of the present discussion. "A" is the active area of the structure, "a" is the distance from the cathode to the plane of the grid, "b" is the distance from the plane of the grid to the plate, E_b is the plate voltage, E_{c1} is the effective grid voltage (including the effect of contact potential) and μ is the amplification constant. Dimensions are in cms. This stipulation is unnecessary for equation (6) but is necessary for some of the later equations. For a plane parallel tetrode or pentode, equation (6) is a good approximation if E_b is replaced by the screen voltage E_{c2}, μ is replaced by the "triode mu" μ_{12} (*mu* of control grid with respect to screen grid) and the coefficient M introduced, where M is the ratio of the plate current to the cathode current. The reason this approximation is good is that the field at the cathode, and therefore the cathode current, is determined almost entirely by the potentials of the first two grids. Because of the screening effect of these grids the potential of the plate has little effect on the cathode current. We have then

$$I_b = \frac{2.33 \times 10^{-6} MA \left(E_{c1} + \dfrac{E_{c2}}{\mu_{12}}\right)^{3/2}}{a^2 \left(1 + \dfrac{1}{\mu_{12}}\dfrac{a+b}{a}\right)^{3/2}} \quad \dots\dots\dots\dots(7)$$

The distance "b" is now the distance from the plane of the grid to the plane of the screen. This expression can be differentiated with respect to E_{c1} to get

$$G_m = \frac{dI_b}{dE_{c1}} = \frac{3}{2} \frac{2.33 \times 10^{-6} MA \left(E_{c1} + \dfrac{E_{c2}}{\mu_{12}}\right)^{1/2}}{a^2 \left(1 + \dfrac{1}{\mu_{12}}\dfrac{a+b}{a}\right)^{3/2}} \quad \dots\dots(8)$$

It is assumed that μ_{12} and M are independent of E_{c1}. If we let I_0 be the cathode current density, then substitute $MI_0 A$ for I_b in (7), and eliminate the expression $\left(E_{c1} + \dfrac{E_{c2}}{\mu_{12}}\right)$ between (7) and (8), we have

$$G_m = \frac{3}{2} \frac{(2.33 \times 10^{-6})^{2/3} MA I_0^{1/3}}{a^{4/3} \left(1 + \dfrac{1}{\mu_{12}}\dfrac{a+b}{a}\right)} \quad \dots\dots\dots\dots(9)$$

Neglecting the stray capacitances between the lead wires and also the edge effects, the greater proportion of the cold capacitance (input plus output) in Farads can be approximated by

$$C_0 = .0885A \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} \right) \times 10^{-12} \cdot \cdots \cdots \cdots (10)$$

where "a" and "b" have the same meaning as in (7) and "c" is the spacing between the plate and the suppressor (or screen in the case of a tetrode). This is of course a highly idealized case. The cathode, the grids, and the plate, are each assumed to be plane conductors of infinitesimal thickness, each having an area equal to the active area of the structure.* The band merit then becomes.

$$B_0 = \frac{G_m}{2\pi C_0} = \frac{4.74 \, M \, I_0^{1/3} \times 10^8}{a^{4/3} \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} \right) \left(1 + \frac{1}{\mu_{12}} \frac{a+b}{a} \right)} \quad \cdots \cdots (11)$$

With a given cathode current, the factor M increases as the screen current is reduced. The use of small wires for the screen grid is an important factor in obtaining minimum screen current.

I_0, the useful cathode current density, is limited by the emission capabilities of the cathode. In practice it is necessary to operate in a region considerably below the maximum available emission to avoid excessive changes in transconductance which would result from variations in cathode activity with time. Also, the shot noise will begin to rise when the region of temperature-limited operation is approached. It should be noted that B_0 is independent of the area A.

Taking reasonable values for M, I_0 and μ_{12}, ($M = 0.75$, $I_0 = 50 \, ma/cm^2$, $\mu_{12} = 25$), equation (11) becomes

$$B_0 = \frac{1.31 \times 10^8}{a^{4/3} \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} \right) \left(1 + .04 \frac{a+b}{a} \right)} \cdot \cdots \cdots (12)$$

The curves in Figs. 6, 7 and 8 show how B_0 varies with each of the variables "a", "b" and "c" when a constant value is assigned to the other two. The band merit shown on these curves is considerably greater than that which can be realized in an actual circuit of the simple band-pass type assumed because the stray capacitances in the tube, the socket, the circuit elements, and the wiring increase the total interstage capacitance substantially. Also, the grid-cathode capacitance is substantially higher under normal operating

---

* This assumption is obviously not true, particularly in the case of the suppressor grid, but, by simply regarding the effective suppressor-plate spacing as somewhat greater than the actual spacing, the assumption becomes useful.

conditions, because of the presence of space charge, than when the tube is cold, as assumed in deriving equation (11). It has been assumed in making
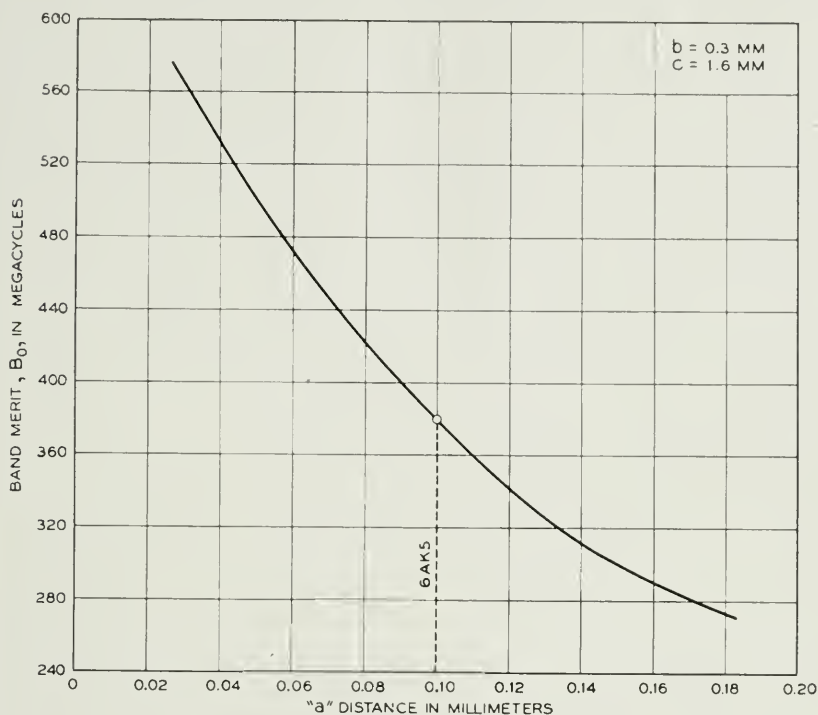
Fig. 6—Band merit vs. grid-cathode spacing.
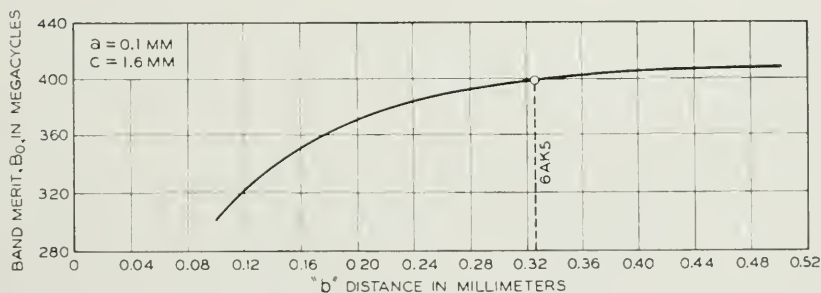
Fig. 7—Band merit vs. grid-screen spacing.

the calculations that the current density and the triode *mu* are held constant while "*a*", "*b*" and "*c*" are varied. In the cases of "*a*" and "*b*", this requires that the control grid and/or screen voltage be varied in such a fashion as to hold I_0 constant. The current density is nearly independent of "*c*" over a

reasonable range provided "c" is not so large as to cause the formation of a virtual cathode in the screen-plate space.

Figure 6 shows that B_0 rises as "a" is reduced.   Reducing "a" also has the advantage that a lower screen voltage is required with a given grid bias. The improvement in B_0 as "a" is reduced is quite rapid, but of course the mechanical difficulties involved in reducing "a" below about 0.10 mm become very great.

The curve in Fig. 7 shows that B_0 does not increase much if "b" is increased above about 0.30 mm, and increasing "b" has the disadvantage that higher screen voltage for a given grid bias is required.

The curve in Fig. 8 shows that B_0 increases very slowly if "c" is increased beyond about 1.50 mm, and increasing "c" has the disadvantage that the external dimensions of the structure become greater.   Increasing "c" also means that the plate comes closer to any shield which is placed around the
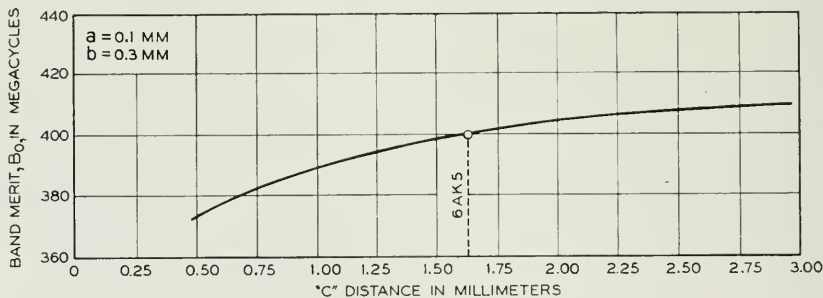
Fig. 8—Band merit vs. plate-suppressor spacing.

bulb of the tube, thus increasing the output capacitance and tending to cancel out any improvement obtained in the structure itself.   Also, if "c" is made too large a virtual cathode may be formed in the screen-plate space under some conditions.   This would interfere with the normal operation of the tube.

In order to take full advantage of the close grid-cathode spacing, the control-grid pitch should be no greater than about 1.5 times the spacing and the grid wires should be as small as possible.   If the pitch is too large, the parts of the cathode directly opposite the grid wires will be cut off while space current is flowing from the sections opposite the spaces between grid wires.   This state of affairs shows up in the characteristics as excessive variation of the triode amplification factor, μ_{12}, as the grid bias is varied, and results in a reduction of the transconductance.   That is, when the grid is made more negative the amplification factor μ_{12} decreases so that the plate current does not decrease as much as it would if μ_{12} remained constant.

When the grid is made more positive the plate current rise is reduced because μ_{12} increases. The grid wires should be as small as possible so as to block off no more of the area of the structure than is necessary. An ideal grid would be an infinitesimally thin conducting plane which offered no resistance to the passage of electrons through it except that due to its electrostatic potential (sometimes called a "physicist's grid"). In the 6AK5 tube the grid-cathode clearance is 0.089 mm (0.0035 inch), the control-grid pitch is 0.0127 mm (0.0050 inch), and the wire size is 0.0010 inch. Experiments have shown that substantially higher transconductance could have been realized, with the same spacing, if smaller wires and smaller pitch had been used, but the mechanical difficulties would have been much greater.

The way to achieve a high band merit from the tube design standpoint is thus to use as close grid-cathode clearance as practicable, to operate the tube at as high a current density as the emission capabilities of the cathode will permit, and to keep the stray capacitances as low as possible. It was noted above from (11) that B_0 is independent of A. However, if it were possible to maintain the same grid-cathode clearance with a large tube as it is with a small one, the larger tube would have the advantage that the stray capacitances in the tube would be a smaller fraction of the total capacitance so that the band merit for the tube would be higher. It would also be closer to what can be realized when the tube is used in an actual circuit because the capacitances added by the socket, the wiring and the circuit elements would be less important. However, the practical mechanical limitations controlling the minimum grid-cathode clearance have been such that the band merit is roughly independent of the tube size over a moderate range of sizes of high transconductance receiving tubes.

3. Input Conductance

Two factors tend to make the input conductance of tubes higher at high frequencies than at low frequencies. One is the effect of lead inductances and the other is the effect of transit time. If the loading produced by these effects is no more than that required to get the desired bandwidth, it may be no particular disadvantage for stages other than the first one in the amplifier. As will be seen later, however, this effect in the input tube increases the noise figure. The practical result of a consideration of these effects is that the leads are made as short as possible and that small tubes are used in order to use close grid-cathode clearances when the frequency at which the tubes are to be used is above about 10 mc. The expression derived by North[3] for

---

[3] "Analysis of the Effects of Space Charge on Grid Impedance," D. O. North, I. R. E. Proceedings, Vol. 24, No. 1, January, 1936.

the input conductance of a tetrode or pentode can be re-written, neglecting the higher order terms, to give the approximate expression

$$G_{in} = \frac{5.0 \times 10^{-3} a^2 f^2 G_m}{V_1} \left[1 + 3.3 \frac{b}{a \left(1 + \sqrt{\frac{V_2}{V_1}} \right)} \right] \quad \ldots (13)$$

where G_m is the triode-connected transconductance, a is grid-cathode spacing in cms, b is grid-screen spacing in cms, V_1 and V_2 are the effective grid-plane and screen-plane potentials in volts, and f is the frequency in mc. It is assumed that there are no lead inductances and that there is no potential minimum in the grid-cathode region. For a given transconductance, (13) shows that close spacings are necessary for minimum grid conductance.

If the lead inductances between the external circuit and the tube elements are appreciable, the input loading may be excessive even though the transit time through the tube structure is negligibly small. The general case taking account of the mutual and self inductances of all the leads of a pentode has been treated by Strutt and van der Ziel[4]. The equations are cumbersome even though only the first order terms in frequency are retained. If all of the lead inductances except that in series with the cathode are neglected, and transit time is assumed to be negligible, the input conductance of a pentode becomes approximately[5]

$$G_{in} = \omega^2 G_m L_k C_k. \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots (14)$$

where L_k is the cathode lead inductance and C_k is the grid-cathode capacitance. It is further assumed that the plate-grid capacitance is negligible.

Work done at the Naval Research Laboratory includes data on the input conductance of 6AK5 tubes in the frequency range from 100 mc to 300 mc. Through the courtesy of the Naval Research Laboratory some of the data are reproduced in Fig. 9. It is of interest to check a point on this curve against equation (14). For the 6AK5, 0.02 micro-henry is the estimated* cathode lead inductance, $G_m = 5000 \times 10^{-6}$ mhos, and $C_k = 4 \times 10^{-12}$ farad. At a frequency of 250 mc we have a calculated conductance of 990 micromhos, which checks roughly with the value of 1110 micromhos from the curve in Fig. 9. Equation (13) can be used to obtain an approximate value for the loading due to transit time. Taking $a = 0.0089$ cm., $b = 0.032$ cm., $G_m = 6.7 \times 10^{-3}$ mhos, $f = 250$ mc, $V_1 = +2.3$ volts, and $V_2 = +120$ volts, a calculation gives $G_{in} = 177$ micromhos. These results

---

[4] "The Causes for the Increase of the Admittance of Modern High-Frequency Amplifier Tubes," M. J. O. Strutt and A. van der Ziel, *I. R. E. Proceedings*, Vol. 26, No. 8, August 1938.

[5] "Hyper and Ultra-High Frequency Engineering," Sarbacher and Edson, p. 435.

* This has been checked roughly by Q-meter measurements.

show that the performance of the 6AK5 near the upper end of its present useful frequency range is probably limited to a considerable degree by the lead inductances rather than by transit time effects in the structure.
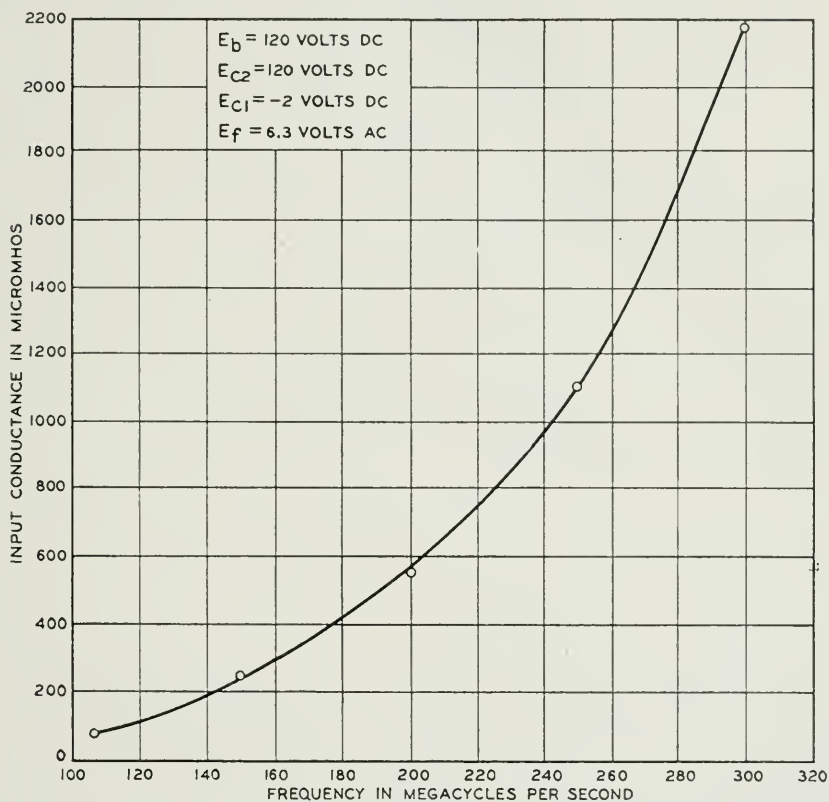
Fig. 9—Average input conductance vs. frequency for six 6AK5 tubes—courtesy of Naval Research Laboratory.

4. Noise

Although it may be possible to employ enough stages of IF amplification to provide the necessary gain and band-width we may still have a relatively insensitive receiver for weak signals. This comes about because there are inherent electrical disturbances in vacuum tubes and passive networks which give rise to random voltages. Since these disturbances may be of the order of the strength of the signal, they must be kept to a minimum in order to maintain a high signal-to-noise ratio. In a receiving system in which no RF amplification is used ahead of the first detector, the signal-to-noise ratio is limited to a large extent by the noisiness of the first detector and the first

IF tube. The noise performance of the first IF stage will be discussed in some detail.**

A convenient method of expressing the departure from ideal performance is the use of the "noise figure" proposed by Friis.[6] The noise figure of a network or amplifier may be defined as follows:

$$NF = \frac{\text{Available output noise power}}{GKT\Delta f} \dots\dots\dots\dots (15)$$

where G is the "available gain" which is defined as the ratio of the available signal power at the output terminals of the network or amplifier to the available signal power at the terminals of the signal generator. $KT\Delta f$ is the available noise power from a passive resistance, where K is Boltzmann's constant, T is absolute temperature and Δf is the incremental bandwidth. This follows from consideration of a noise generator of resistance R_0 working into a load resistance R_0. This is the condition for maximum power into the load, half of the noise voltage appearing across the source and half across the load. The open-circuit noise voltage appearing across the terminals of a resistance R_0 is

$$V^2 = 4KTR_0\Delta f \dots\dots\dots\dots\dots (16)$$

The noise power delivered to the load by the source resistance will be

$$P_n = \frac{V^2}{4R_0} = KT\Delta f \dots\dots\dots\dots\dots (17)$$

If there were no source of noise other than that of the resistance of the signal source itself, the noise figure would be unity. If the signal source works directly into a matched load resistance at the same temperature, the noise figure is 2 since the available output noise power is $KT\Delta f$ and the available gain is one-half.

The importance of the noise figure of the radar receiver is obvious since a reduction in the noise figure is equivalent to the same percentage increase in transmitter power. In recent radar systems the noise arising in the first IF stage constituted a substantial part of the total receiver noise.

If the first IF tube provides at least 15 db gain, noise introduced by its plate load impedance, and by any other sources in the rest of the amplifier, is usually negligible. The departure of the noise figure of the IF amplifier from unity is then due to noise arising in the first tube and in its input circuit.

In the very high frequency (VHF) range, essentially all of the noise arising

---

** It will be assumed in all of the discussion about noise that there is no noise due to flicker effect, emission of positive ions, microphonics, sputter, ionization, secondary emission, or reflection of electrons from charged insulators.

[6] "Noise Figure of Radio Receivers," H. T. Friis, *I. R. E. Proceedings*, July 1944.

in the tube itself is due to random fluctuations in the emission of electrons from the cathode. For a parallel plane diode in a circuit such as that shown in Fig. 10(a), the mean square noise current is[7]

$$i_k^2 = 2\Gamma^2 e I_k \Delta f \dots\dots\dots\dots\dots\dots\dots(18)$$

where e is the electronic charge, I_k is the d-c cathode current, Δf is the incremental bandwidth, and Γ is a factor which takes into account the

(a)

(b)

(c)
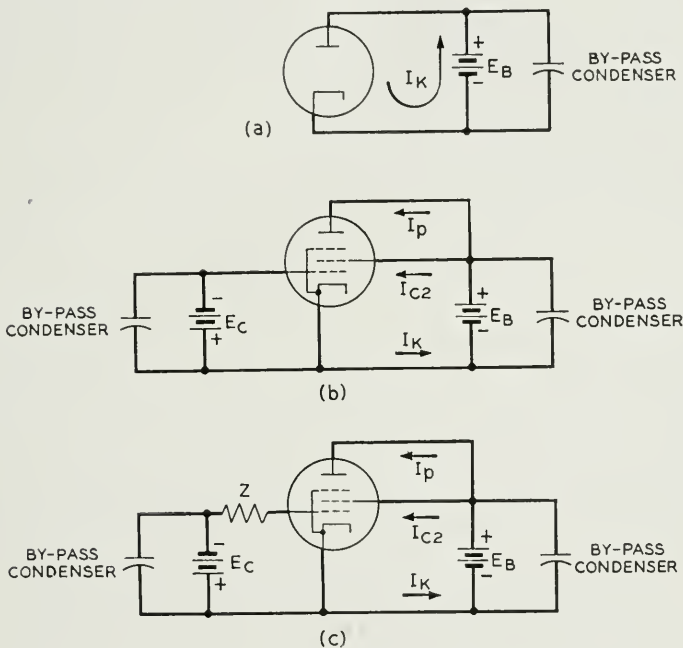
Fig. 10a—Diode, negligible circuit impedance.
Fig. 10b—Pentode, negligible circuit impedance.
Fig. 10c—Pentode, impedance in grid circuit.

"cushioning effect" of space charge. For anode-cathode spacings and operating conditions such that the transit time is not too large a faction of the period of the frequency involved, Γ is of the order of 0.20 when the zero-field emission is several times larger than I_k. Under temperature-limited conditions Γ is unity. That is, under favorable space-charge-limited conditions, the fluctuation noise component in I_k is only about 20% of the value for the same cathode current under temperature-limited conditions. Although the introduction of grids between the cathode and anode

[7] "Fluctuations in Space-Charge-Limited Currents at Moderately High Frequencies," B. J. Thompson, D. O. North, W. A. Harris, *R. C. A. Review*, October 1940, Vol. 5, p. 244.

of the simple diode complicates the noise problem, the factor Γ remains of great importance. It is therefore desirable to control the tube processing in manufacture so as to insure adequate available emission in every tube. An "activity test" is made on completed tubes for this purpose. This test consists of reducing the heater voltage by an arbitrary amount (usually 10%) and observing the change in one of the tube characteristics which is sensitive to changes in available emission. The characteristic used for tubes like the 717A and 6AK5 is the transconductance. If the transconductance decreases by more than about 20% for a 10% reduction in heater voltage, with the other operating voltages held constant, insufficient available emission is indicated and Γ is higher than for more "active" tubes.

In the case of a pentode with negligible impedance in each of its leads, as shown in Fig. 10(b), the fluctuation in the cathode current is the same as that given in equation (18), but the noise component in the plate lead is larger. Thompson, North, and Harris[7] showed that it can be written as

$$i_p^2 = 2eI_p\Delta f\left[\frac{\Gamma^2 I_p + I_{c2}}{I_k}\right] \quad\ldots\ldots\ldots\ldots\ldots(19)$$

where I_p is the d-c plate current and I_{c2} is the d-c screen current. It was mentioned above that Γ can be made as low as 0.20 by providing adequate available emission. From the design standpoint, equation (19) also shows that the screen current should be as small as possible. For normal operating conditions in a pentode, the screen current is influenced by the screening fraction (fraction of area blocked off by grid wires) of the screen grid and by the amount of space current turned back to the screen in the screen-plate region. The way to get a minimum screening fraction and still obtain the desired function of the screen grid of reducing the plate-grid capacitance is to use wire of as small diameter as possible. The presence of a suppressor grid, at cathode potential, placed between the screen and the plate to prevent interchange of secondary electrons, causes a certain proportion of the space current which would otherwise go to the plate to be turned back to the screen. Here again, this effect is minimized by using as fine wire as possible for the suppressor grid. It was pointed out in an earlier section that fine wires are desired for the control grid. The ideal for each of the three grids in an IF pentode would be a conducting plane which offers no resistance to the passage of electrons other than the influence of its potential.

When impedances are connected in the various leads of a pentode the noise components discussed above will, in general, be different due to the influence of fluctuation voltages developed between the elements of the tube. In particular it is found experimentally in the VHF range of frequencies that when an impedance Z is introduced between the grid and cathode, as shown in Fig. 10(c), the noise component in the plate lead rises more than

would be expected due to thermal noise from Z, because of the effect of grid noise. North and Ferris[8] showed that the grid noise can be taken into account by assuming that the input resistance of the tube is a resistance noise source whose absolute temperature is about 4.8 times ambient, if the input loading is due to transit time effects alone. One consequence of this input loading is that at high frequencies the best signal-to-noise ratio is usually obtained with an input circuit of lower impedance than that which would be used at low frequencies.

Actually, as was brought out in an earlier section, the loading in tubes like the 6AK5 is probably due largely to lead inductances between the active tube elements and the external circuit components up to a few hundred megacycles. According to Pierce[9] the effect of the cathode lead inductance feedback is to reduce the signal component in the output current while leaving the noise current due to screen interception noise unaffected. Input loading may be a limiting factor in tubes like the 717A and 6AK5 when the frequency is of the order of 100 mc or higher, both because of its effect on gain in some cases and because of its adverse effect on the signal-to-noise ratio for early stage use. There is good evidence that the 6AK5 structure would be useful at much higher frequencies than is the case at present if the circuit connections to the tube elements were improved by more advantageous mounting of the structure and the use of more suitable sockets or external connectors.

Noise measurements made by a number of workers at Bell Telephone Laboratories[10] indicate that an average noise figure of about 2.8 can be obtained with the 6AK5 at a midband frequency of 60 mc, with a well-designed input circuit, and bandwidths up to 10 mc. At a mid-band frequency of 30 mc the noise figure is about 2.4. At 100 mc it is about 3.6.

5. DESCRIPTION OF THE DESIGN OF THE WESTERN ELECTRIC 6AK5 TUBE

In order that the reader may have a full appreciation of the dimensions and other requirements of design to meet the characteristics discussed above, a detailed description of the 6AK5 tube follows.

5.1 Mechanical Description

The 6AK5 tube is an indirectly heated cathode type pentode employing the 7-pin button stem and the T-5-1/2 size miniature bulb. The outline dimensions are shown in Fig. 11. A photograph of a mount ready to be sealed into a bulb is shown in Fig. 12. Figure 13 is a photograph of a transverse section through the tube at the middle of the structure in a plane

[8] "Fluctuations Induced in Vacuum Tube Grids at High Frequencies," D. O. North and W. R. Ferris, *I. R. E. Proceedings*, Vol. 29, No. 2, February 1941.
[9] Unpublished Technical Memorandum, J. R. Pierce.
[10] S. E. Miller, V. C. Rideout, R. S. Julian.
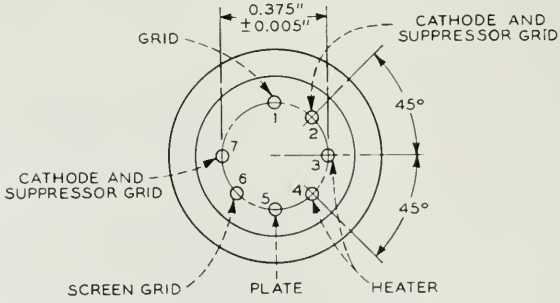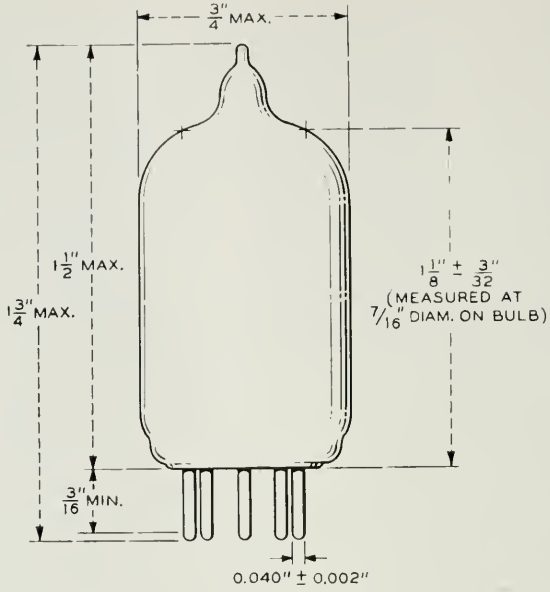
Fig. 11—6AK5 outline dimensions.
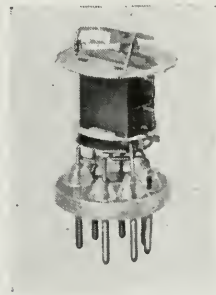
Fig. 12—6AK5 mount structure (full size).

parallel to the stem or "base" of the tube.   The magnification in Fig. 13 as reproduced is about 5½ times.

The heater is a conventional folded type with eight legs of coated tungsten wire.   The wire diameter is 0.0014 inches and the unfolded length is 3 inches. The insulating coating consists of fired aluminum oxide and is about 0.0025 inch thick.

The cathode is of oval cross-section with the contours of the longer sides shaped to conform to the shape of the control grid.   The major axis of the
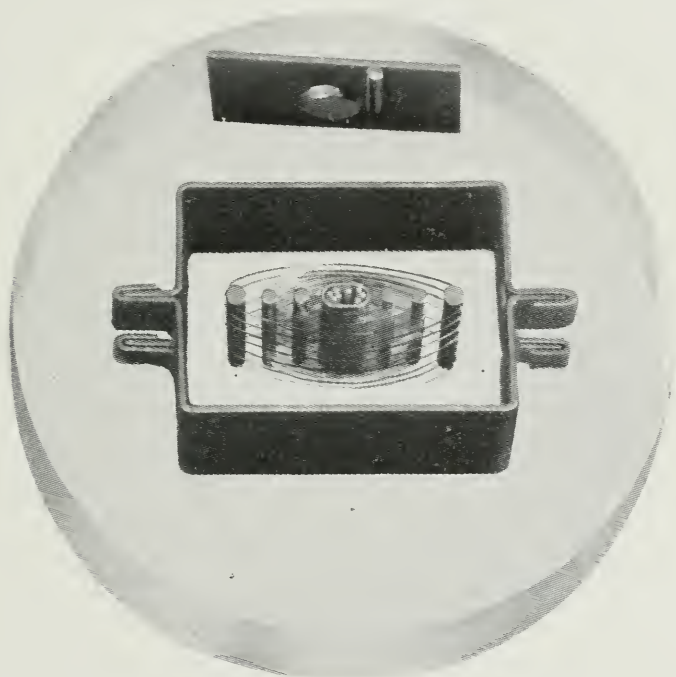
Fig. 13—Transverse section of 6AK5 tube (5½ times).

cross-section of the cathode sleeve is 0.048 inch.   The minor axis is 0.025 inch.   The length of the sleeve is 0.47 inch and it is coated over a centralized section which extends 0.28 inch along its length.   The coating is the usual mixture of oxides of barium, calcium, and strontium.   Before the tube is processed, the coating thickness is about 0.002 inch.   After processing, it is about 0.001 inch thick.

The grids are oval-shaped and are wound with small diameter wires. The control-grid and the screen-grid lateral wires are 0.001 inch diameter tungsten.   The suppressor-grid lateral wires are 0.002 inch diameter molybdenum.   The control-grid wires are gold-plated in order to minimize primary emission of electrons from this grid.

The plate has a rectangular transverse cross-section and is made of 0.005 inch thickness carbonized nickel. The carbonization increases the thermal emissivity of the plate surface so that a reasonable amount of power dissipation in the plate can be tolerated.

The end shields are made of nickel-plated iron. The reason for using this material instead of nickel, as is more often the case, is to prevent overheating of these shields during the exhaust process. At the temperatures used, the iron shields pick up less energy in the induction field during the out-gassing of the plate than do nickel shields. The function of the end shields is to minimize the stray capacitance between the plate and the control grid.

The insulators or spacers which hold the tube elements in proper disposition with respect to each other are of high grade mica. They are coated with magnesium oxide to minimize surface leakage effects.

The getter, which can be seen at the top of the structure in Fig. 12, contains barium which is flashed onto the inside surface of the bulb at the end of the exhaust process in order to take up residual gas evolving from the parts of the tube during operation.

Although the individual parts are extremely small and fragile, the completed tube is surprisingly rugged. The short supporting wires in the stem and the support provided by the bulb-contacting top insulator result in the stem, bulb, and mount structure being a relatively rigid unit. The small parts assembled into the mount are very light in weight and therefore exert relatively small forces on their supporting members under conditions of mechanical shock. Shock tests performed at the Naval Research Laboratory and at the Bell Laboratories show that the 6AK5 tube stands up satisfactorily under a steady vibration of rms acceleration 2.5 times gravity and withstands 1 millisecond shocks of over 300 times gravity.

The most important single geometrical factor in the tube is the spacing between the cathode and the control grid. In the 6AK5 tube this clearance is 0.0035 inch after processing. Before processing it is 0.0025 inch. The manufacturing difficulties involved in assembling the structure and maintaining such small clearances are obviously very great. However, as was seen from the discussion above, this close spacing is essential in order to obtain the desired high-frequency performance.

It can be observed that the close mounting of the structure on the stem provides very short lead lengths between the tube elements and the external pins. This is of importance at frequencies where the inductances of the lead wires become comparable in magnitude to the circuit reactances.

5.2 *Low Frequency Electrical Charatceristics*

The usual static characteristics are shown by the curves in Fig. 14. The ratings, nominal characteristics, and cold capacitances are given in Table II

5.3 *Fixed Tuning*

It is highly desirable to design the IF amplifier with fixed tuning in order to minimize the number of adjustments that need to be made when tubes
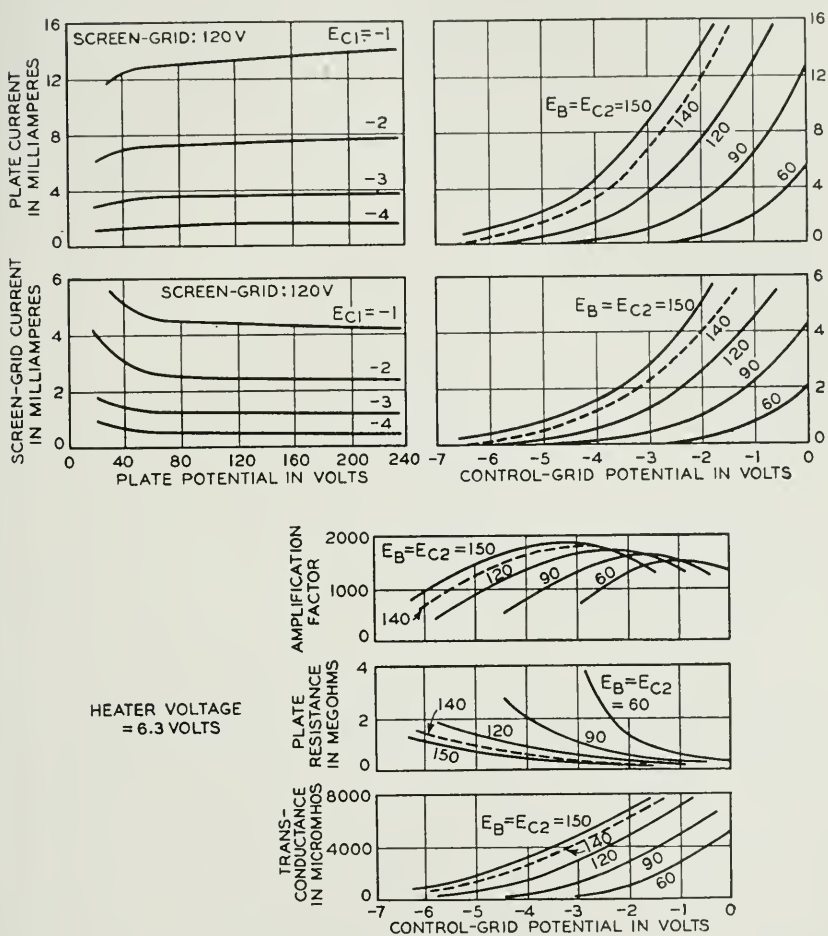
Fig. 14—Average 6AK5 characteristics.

are changed. The difficulty immediately arises that, since the tube capacitances are usually a large fraction of the total shunt capacitance of each coupling network, the variations in capacitance from one tube to another become very important. It is essential that, after the amplifier is lined up at the factory with standard tubes, any stock tubes taken at random shall give satisfactory performance in the amplifier. This requires close control

of the capacitances. Not only must the capacitances be held within the limits of the test specification, but the product averages must be kept close to the design center values particularly in the cases of the input and output capacitances.

TABLE II

HEATER RATING

Heater voltage	6.3 volts, a-c or d-c
Nominal heater current	0.175 ampere

Maximum Ratings (Design-center values)

Maximum plate voltage	180 volts
Maximum screen voltage	140 volts
Maximum plate dissipation	1.7 watts
Maximum screen dissipation	0.5 watt
Maximum cathode current	18 milliamperes
Maximum heater-cathode voltage	90 volts
Maximum bulb temperature	120°C

OPERATING CONDITIONS AND CHARACTERISTICS

Plate voltage	120	150	180 volts
Screen voltage	120	140	120 volts
Cathode-bias resistor	200	330	200 ohms
Plate current	7.5	7.0	7.7 milliamperes
Screen current	2.5	2.2	2.4 milliamperes
Amplification factor	1700	1800	3500
Plate resistance	0.34	0.42	0.69 megohms
Transconductance	5000	4300	5100 micromhos

INTERELECTRODE CAPACITANCES (*With JAN 1A No. 314 shield connected to cathode*)

Control grid to heater, cathode, screen grid and suppressor grid	3.90 $\mu\mu f$
Plate to control grid	0.01 $\mu\mu f$
Plate to heater, cathode, screen grid and suppressor grid	2.85 $\mu\mu f$

6. CONCLUSION

The important factors to be considered in evaluating the suitability of a vacuum tube for broad band IF use in the VHF range are as follows:

1. Band merit
2. Noise figure
3. Input conductance
4. Power consumption
5. Physical size
6. Control of capacitances

We have seen that the tube design features which make a tube good on the basis of these requirements are close grid-cathode spacing, fine grid wires, short lead wires, and small elements. An important consideration from the manufacturing standpoint is the control of cathode emission for low noise. It has been possible to extend the useful frequency range of conventional receiving tubes up through the VHF range and somewhat higher by this line of attack.

The 6AK5 tube is an outgrowth of the many years of development in this general field by the Electronics group of the Bell Telephone Laboratories. Contributions to the success of the development have been made by chemists, physicists, mechanical engineers and electrical engineers too numerous to mention individually and to whom the author is indebted for much of the material presented.

High Q Resonant Cavities for Microwave Testing

By I. G. WILSON, C. W. SCHRAMM and J. P. KINZER

Formulas and charts are given which aid the design of right circular cylinder cavity resonators operating in the $TE\ 01n$ mode, which yields the highest Q for a given volume. The application of these to the design of an echo box radar test set is shown, and practical considerations arising in the construction of a tunable cavity are discussed.

INTRODUCTION

A TUNABLE high Q resonant cavity is a particularly useful tool for determining the over-all performance of a radar quickly and easily[1].* Further, since it uses the radar transmitter as its only source of power, it can be made quite portable. When a high Q cavity is provided with two couplings, one for the radar pickup and the other to an attenuating device, crystal rectifier and meter which serve for tuning the cavity not only can an indication of over-all performance be obtained but other useful information as well. For example, the transmitter frequency can be measured; calibration of the crystal affords a rough measure of the transmitter power; and an analysis of the spectrum can be made by plotting frequency versus crystal current. This information is of particular importance in radar maintenance.

The Q required for this purpose is quite high, comparable to that obtained from quartz crystals in the video range. For this reason, such cavities have many additional possibilities for use in microwave testing equipment and microwave systems. For example, they may form component parts of a narrow band filter, or be used as discriminators for an oscillator frequency control.

Resonant cavities are of two general types—tuned and untuned. A tuned cavity is designed to resonate in a single mode adjustable over the radar frequency range. An untuned cavity is of a size sufficient to support a very large number of modes within the working range. Both are useful, but the tuned variety can give more information about the radar and hence has been more widely used.

While a tuned cavity may be a cylinder, parallelepiped or sphere (or even other shapes), the first of these has been most thoroughly explored by us. It offers the possibility of utilizing the anomalous circular mode, described by Southworth[2] in his work on wave guides, which permits the attainment of high Q's in quite a small size. In addition, it is easier to construct a variable length cylinder than a variable sized sphere.

* Superscripts refer to bibliography.

Due to their interesting properties the history of resonators of the cavity type in which a dielectric space is enclosed by a conducting material, goes back many years. In 1893 J. J. Thompson[3] derived expressions for resonant frequencies of the transverse electric modes in a cylinder. Lord Rayleigh[4] published a paper in 1897 dealing with such resonant modes. The early work was almost entirely theoretical but some experiments were carried out in 1902 by Becker[5] at 5 and 10 centimeters. In recent years, the subject has been fairly thoroughly investigated (at least theoretically) for several simple shapes.

However, many of the presentations are highly mathematical with considerable space devoted to proofs; the results which would be most useful to an engineer are thus sometimes obscured. The purpose of this paper is to present certain engineering results together with information upon the application of the tunable cylindrical cavity to radar testing.

Definitions and Fundamental Formulas

Modes

By fundamental and general considerations, every cavity resonator, regardless of its shape, has a series of resonant frequencies, infinite in number and more closely spaced as the frequency increases. The total number N of these having a resonant frequency less than f is given approximately by:[6]

$$N = \frac{8\pi}{3c^3} V f^3 \qquad (1)$$

in which

V = volume of cavity in cubic meters.

c = velocity of electromagnetic waves in the dielectric in meters per second.

f = frequency in cycles per second.

With each resonance there is associated a particular standing wave pattern of the electromagnetic fields, which is identified by the term "mode."

In right cylinders (ends perpendicular to axis) the modes fall naturally into two groups, the transverse electric (TE) and the transverse magnetic (TM). In the TE modes, the electric lines everywhere lie in planes perpendicular to the cylinder axis, and in the TM modes, the magnetic lines so lie. Further identification of a specific mode is accomplished by the use of indices.

The MS Factor

With the cylinder further restricted to a loss-free dielectric and a non-magnetic surface, there is associated with each mode a value of Q (quality factor)[7] which depends on the conductivity of the metallic surface, on the

frequency and on the shape of the cylinder, e.g. whether it is circular or elliptical, and whether it is slender or stubby. The quantity $\dfrac{Q\delta}{\lambda}$, however, depends only upon the mode and shape of the cylinder and has been referred to as the mode-shape (MS) factor. In this formula, δ refers to skin depth as customarily defined[8], and λ is wavelength in the dielectric, as given by $\lambda = \dfrac{c}{f}$; both δ and λ are in meters.

Fundamental Formulas

Expressions for standing wave patterns and $Q\dfrac{\delta}{\lambda}$ are given in Table I, for right rectangular, circular and full coaxial cylinders*. The table is virtually self-explanatory, but a few remarks on mode designation are needed. The mode indices are l, m, n following the notation of Barrow and Mieher.[9] In the rectangular prism they denote the number of half-wavelengths along the coordinate axes. For the other two cases they have an analogous physical significance with l related to the angular coordinate, m to the radial and n to the axial.

In the elliptical cylinder, a further index is needed to distinguish between modes which differ only in their orientation with respect to the major and minor axes; these paired modes are termed even and odd, and have slightly different resonant frequencies.[10] In the circular cylinder they have the same frequency, a condition which is referred to as a degeneracy (in this case, double); that is, in the circular cylinder, odd and even modes are distinguishable only by a difference in their orientation within the cylinder with reference to the origin of the angular coordinate. In Table I, the field expressions are given for the even modes; those for the odd modes are obtained by changing $\cos l\theta$ to $\sin l\theta$ and $\sin l\theta$ to $\cos l\theta$ everywhere.

The value of N in the table is based on counting this degeneracy as a single mode; counting even and odd modes as distinct will nearly double the value of N, thus bringing it into agreement with the general equation (1). The distinction between even and odd modes is of limited importance in practical applications, and will not be further mentioned.

In Table I, the *mks* system of units is implied. The notation is in general accordance with that used in prior developments of the subject. For engineering applications, it is advantageous to reduce the results to units in ordinary use and to change the notation wherever this leads to a more obvious association of ideas. For these reasons, in what follows attention

---

* The elliptic cylinder (closely allied to the circular cylinder of which it is a generalization) is omitted as the necessary functions are not widely known or easily available.

is confined to the circular cylinder, with changes in units and notation as specified later.

Fig. 1—Mode chart for right circular cylinder resonant cavity.

The Mode Chart

The formula relating the resonant frequency to the mode, shape and dimensions of the cylinder is of prime interest. It may be simply written as

$$(fD)^2 = A + Bn^2 \left(\frac{D}{L}\right)^2 \tag{2}$$

where

f = frequency in megacycles per second.

D = diameter of cavity in inches.

L = length of cavity in inches.

A = a constant depending upon the mode. Values of A are given in Table II for the lowest 30 modes. Values of Bessel function roots are given in Table III for the first 180 modes.

B = a constant depending upon the velocity of electromagnetic waves in the dielectric. For air at 25°C and 60% relative humidity, $B = 0.34799 \times 10^8$.

n = third index defining the mode, i.e., the number of half wavelengths along the cylinder axis.

TABLE II.—*Constants for Use in Computing the Resonant Frequencies of Circular Cylinders*

$$(fD)^2 = \left(\frac{cr}{\pi}\right)^2 + \left(\frac{cn}{2}\right)^2\left(\frac{D}{L}\right)^2 = A + B\,n^2\left(\frac{D}{L}\right)^2$$

$B = 0.34799 \times 10^8$ $c = 1.17981 \times 10^{10}$ inches/second

Mode	r	A
TM 01	2.40483	0.81563×10^8
02	5.52008	4.2975
03	8.65373	10.5617
11	3.83171	2.0707
12	7.01559	6.9415
13	10.17347	14.5970
21	5.13562	3.7197
22	8.41724	9.9923
31	6.38016	5.7410
32	9.76102	13.4374
41	7.58834	8.1212
51	8.77148	10.8511
61	9.93611	13.9238
TE 01	3.83171	2.0707
02	7.01559	6.9415
03	10.17347	14.5970
11	1.84118	0.47810
12	5.33144	4.0088
13	8.53632	10.2770
21	3.05424	1.3156
22	6.70613	6.3426
23	9.96947	14.0175
31	4.20119	2.4893
32	8.01524	9.0606
41	5.31755	3.9879
42	9.28240	12.1520
51	6.41562	5.8050
61	7.50127	7.9359
71	8.57784	10.3772
81	9.64742	13.1265

Value of c is for air at 25°C. and 60% relative humidity. D and L in inches; f in megacycles.

Formula (2) represents a family of straight lines, when $(D/L)^2$ and $(fD)^2$ are used as coordinates, and leads directly to the easily constructed and highly useful "Mode Chart" of Fig. 1.

It will be noted from Table II that the $TE\ 0mn$ and the $TM\ 1mn$ modes have the same frequency of resonance. This is a highly important case of degeneracy. In the design of practical cavities it is necessary to take measures to eliminate this degeneracy, as the TM mode (usually referred to as the companion of its associated TE mode) introduces undesirable effects. This is discussed more at length later.

Choice of Operating Mode

Turning now to the expressions for $Q\frac{\delta}{\lambda}$ these are seen to be of a rather complicated nature. For some of the lower order modes, their values are plotted in Figs. 2, 3 and 4. Examination of these leads to the question of which mode has the highest Q for a given volume. It is desirable to keep the volume a minimum, since, as shown by (1), the total number of resonances is a function of the volume. Analysis of the problem is somewhat involved, but leads to the conclusion that operation in the $TE\ 01n$ mode* gives the smallest volume for an assigned Q, and also leads to specific values of n and D/L which give this result. In fact, for maximum Q/V in the $TE\ 01n$ mode,

$$(fD)^2 \left(\frac{D}{L}\right) = 3.11 \times 10^8 \tag{3}$$

which permits easy plotting on a mode chart of the locus of the operating points for best Q/V ratio.

Extraneous or Unwanted Modes

In echo boxes for radar testing, where high Q is of the utmost importance, the $TE\ 01n$ mode has been used. The values of n vary from 1 at frequencies around 1 kmc to 50 at about 25 kmc.

All other modes are then regarded as unwanted or extraneous. The great utility of the mode chart lies in that it permits a quick determination of the most favorable operating area. We consider this now in detail.

Figure 5 shows a portion of a mode chart. It is clear that the sensible way to construct a tunable cylinder is to keep the diameter fixed and vary the length. With fixed diameter, the coordinates of the mode chart are essentially f^2 and $\left(\frac{1}{L}\right)^2$ and it is convenient to refer to them loosely as fre-

---

* Unimportant exceptions occur for values of $Q\frac{\delta}{\lambda} < 1.2$.

quency and length.   With this understanding, if the frequency band to be covered by the tunable cavity extends from f_1 to f_2 then the length must be adjustable from L_1 to L_2.   Responses to frequencies within the band, but
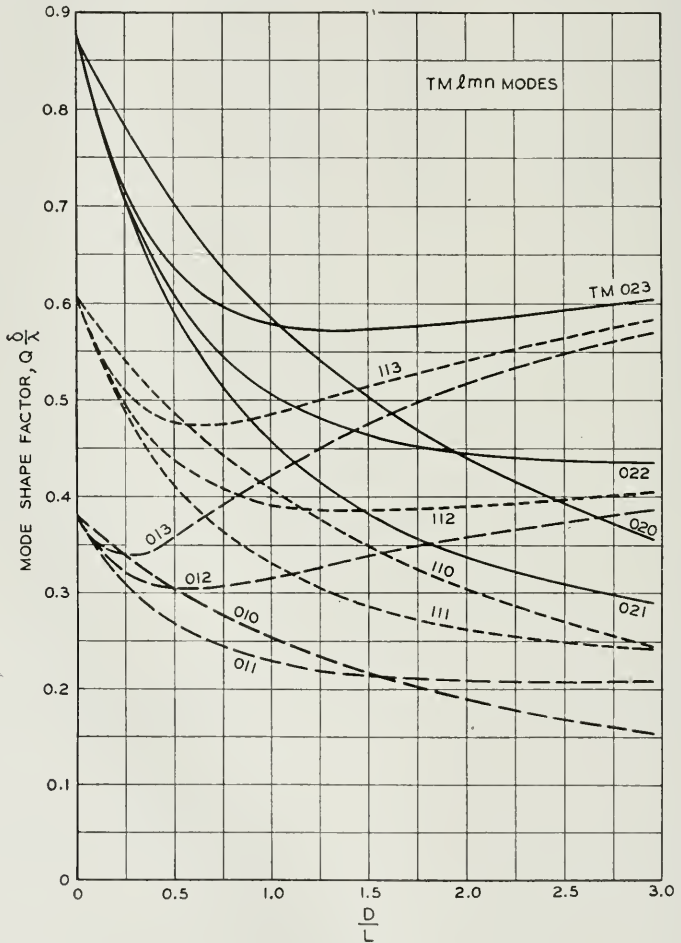
Fig. 2—Mode-shape factors $\left(Q\dfrac{\delta}{\lambda}\right)$ as a function of diameter to length ratio $\left(\dfrac{D}{L}\right)$ for circular cylinder resonator—*TM* modes.

which demand lengths outside the range L_1 to L_2 are of little interest because mechanical stops prohibit other lengths.   On the assumption that the applied frequency will always lie within the operating band f_1 to f_2, responses to frequencies below f_1 or above f_2 are likewise of little interest.   Therefore,

major consideration need be given only to those modes which lie within the
rectangle of which the desired $TE\ 01n$ mode forms the diagonal.
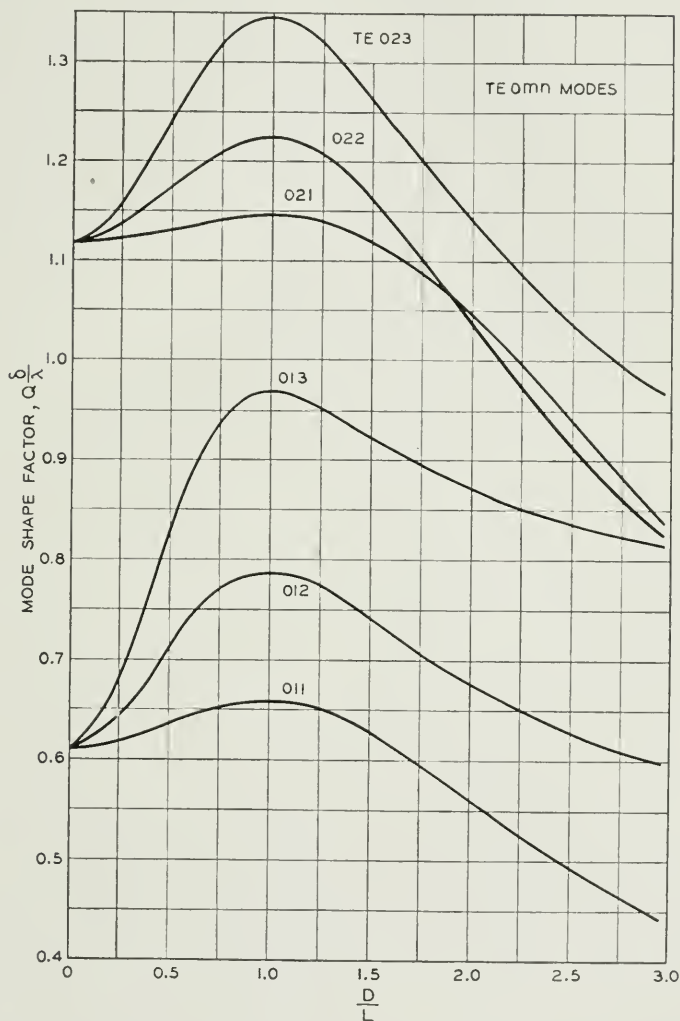
Fig. 3—Mode-shape factors $\left(Q\frac{\delta}{\lambda}\right)$ as a function of diameter to length ratio $\left(\frac{D}{L}\right)$ for circular cylinder resonator — TE modes with $l = 0$.

Any nondiagonal mode is an extraneous mode. Those which do not
cross the desired mode within the frame are called interfering modes. They
act to give responses at more than one tuning point when the applied fre-

quency is held fixed, or alternatively to give responses at more than one frequency when the tuning is held fixed. In either event they lead to ambiguity and confusion, and their effects must be reduced to the point where this cannot occur.
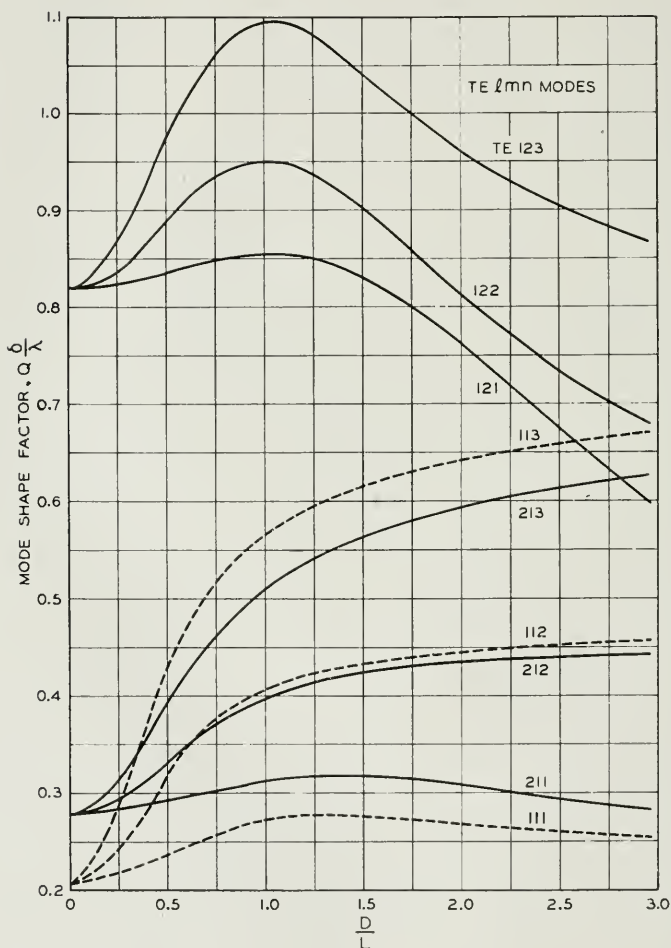
Fig. 4—Mode-shape factors $\left(Q\dfrac{\delta}{\lambda}\right)$ as a function of diameter to length ratio $\left(\dfrac{D}{L}\right)$ for circular cylinder resonator — TE modes with $l > 0$.

A special type of interfering mode is the $TE\,01(n+1)$ mode. In the nature of things, it is virtually impossible to suppress this mode without likewise suppressing the desired $TE\,01n$ mode. The width of the operating

band is thus strictly limited, if ambiguity is to be avoided. This effect, termed self-interference, becomes an important factor as n increases, since

$$\frac{f_2}{f_1} \text{ (maximum)} = \frac{n+1}{n}$$

This maximum value cannot be realized because it is incompatible with other requirements.
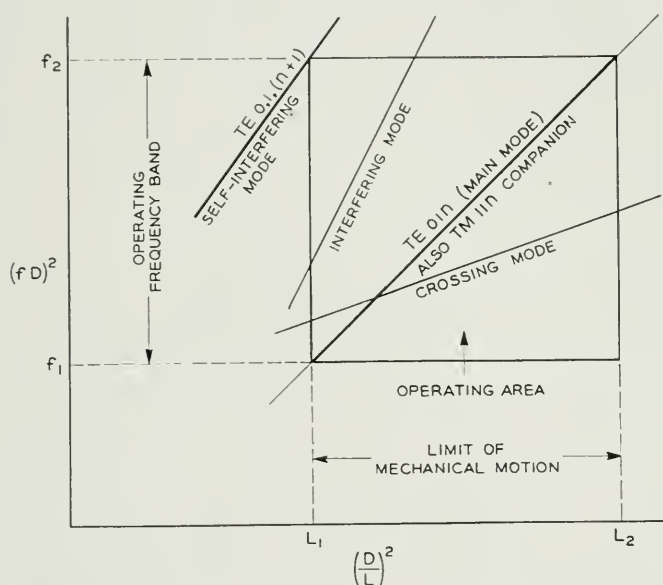
Fig. 5—Mode chart illustrating meaning of crossing and interfering modes and of operating area.

When an undesired mode crosses the main mode within the rectangle it is called a crossing mode. Except in a region close to the crossing point, it acts only to cause ambiguity as already discussed. In the immediate region of the crossing point, however, the cavity is simultaneously resonant in both modes. Violent interaction effects which may seriously degrade the cavity Q frequently occur at such a crossing.

Methods of Minimizing Effects of Unwanted Modes

A major problem in the design of a high Q cavity for radar test purposes, in which the Q and frequency range are set by radar system considerations, is to reduce the effects of all the undesired modes without seriously degrading the main $TE\ 01n$ mode. Among those to be suppressed is the companion $TM\ 11n$ mode which is inherently of the same frequency.

TABLE III.—*Values of the Bessel Function Zero (r_{lm}) for the First 180 Modes in a Circular Cylinder Resonator*

	r_{lm}	Mode*			r_{lm}	Mode*			r_{lm}	Mode*			r_{lm}	Mode*	
1	1.8412	E	1-1	46	13.0152	M	3-3	91	18.4335	M	10-2	136	22.6716	E	2-7
2	2.4048	M	0-1	47	13.1704	E	2-4	92	18.6374	E	6-4	{137	22.7601	M	1-7
3	3.0542	E	2-1	{48	13.3237	M	1-4	93	18.7451	E	12-2	{138	22.7601	E	0-7
{4	3.8317	M	1-1	{49	13.3237	E	0-4	94	18.9000	M	14-1	139	22.9452	M	8-4
{5	3.8317	E	0-1	50	13.3543	M	9-1	95	18.9801	M	5-4	140	23.1158	M	14-2
6	4.2012	E	3-1	51	13.5893	M	6-2	96	19.0046	E	9-3	141	23.2548	E	21-1
7	5.1356	M	2-1	52	13.8788	E	12-1	97	19.1045	E	17-1	142	23.2568	M	18-1
8	5.3176	E	4-1	53	13.9872	E	5-3	98	19.1960	E	4-5	143	23.2643	E	16-2
9	5.3314	E	1-2	54	14.1155	E	8-2	99	19.4094	M	3-5	144	23.2681	E	7-5
10	5.5201	M	0-2	55	14.3725	M	4-3	100	19.5129	E	2-6	145	23.2759	M	11-3
11	6.3802	M	3-1	56	14.4755	M	10-1	101	19.5545	M	8-3	146	23.5861	M	6-5
12	6.4156	E	5-1	57	14.5858	E	3-4	{102	19.6159	M	1-6	147	23.7607	E	10-4
13	6.7061	E	2-2	58	14.7960	M	2-4	{103	19.6159	E	0-6	148	23.8036	E	5-6
{14	7.0156	M	1-2	59	14.8213	M	7-2	104	19.6160	M	11-2	149	23.8194	E	13-3
{15	7.0156	E	0-2	60	14.8636	E	1-5	105	19.8832	E	13-2	150	24.0190	M	4-6
16	7.5013	E	6-1	61	14.9284	E	13-1	106	19.9419	E	7-4	151	24.1449	E	3-7
17	7.5883	M	4-1	62	14.9309	M	0-5	107	19.9944	M	15-1	152	24.2339	M	9-4
18	8.0152	E	3-2	63	15.2682	E	6-3	108	20.1441	E	18-1	153	24.2692	M	15-2
19	8.4172	M	2-2	64	15.2867	E	9-2	109	20.2230	E	10-3	154	24.2701	M	2-7
20	8.5363	E	1-3	65	15.5898	M	11-1	110	20.3208	M	6-4	155	24.2894	E	22-1
21	8.5778	E	7-1	66	15.7002	M	5-3	111	20.5755	E	5-5	156	24.3113	E	1-8
22	8.6537	M	0-3	67	15.9641	E	4-4	112	20.7899	M	12-2	157	24.3382	M	19-1
23	8.7715	M	5-1	68	15.9754	E	14-1	113	20.8070	M	9-3	158	24.3525	M	0-8
24	9.2824	E	4-2	69	16.0378	M	8-2	114	20.8269	M	4-5	159	24.3819	E	17-2
25	9.6474	E	8-1	70	16.2235	M	3-4	115	20.9725	E	3-6	160	24.4949	M	12-3
26	9.7610	M	3-2	71	16.3475	E	2-5	116	21.0154	E	14-2	161	24.5872	E	8-5
27	9.9361	M	6-1	72	16.4479	E	10-2	117	21.0851	M	16-1	162	24.9349	M	7-5
28	9.9695	E	2-3	{73	16.4706	M	1-5	118	21.1170	M	2-6	163	25.0020	E	14-3
{29	10.1735	M	1-3	{74	16.4706	E	0-5	119	21.1644	E	1-7	164	25.0085	E	11-4
{30	10.1735	E	0-3	75	16.5294	E	7-3	120	21.1823	E	19-1	165	25.1839	E	6-6
31	10.5199	E	5-2	76	16.6982	M	12-1	121	21.2116	M	0-7	166	25.3229	E	23-1
32	10.7114	E	9-1	77	17.0038	M	6-3	122	21.2291	E	8-4	167	25.4170	M	16-2
33	11.0647	M	4-2	78	17.0203	E	15-1	123	21.4309	E	11-3	168	25.4171	M	20-1
34	11.0864	M	7-1	79	17.2412	M	9-2	124	21.6416	M	7-4	169	25.4330	M	5-6
35	11.3459	M	3-3	80	17.3128	E	5-4	125	21.9317	E	6-5	170	25.4956	E	18-2
36	11.6198	M	2-3	81	17.6003	E	11-2	126	21.9562	M	13-2	171	25.5094	M	10-4
37	11.7060	E	1-4	82	17.6160	M	4-4	127	22.0470	M	10-3	172	25.5898	E	4-7
38	11.7349	E	6-2	83	17.7740	E	8-3	128	22.1422	E	15-2	173	25.7051	M	13-3
39	11.7709	E	10-1	84	17.7887	E	3-5	129	22.1725	M	17-1	174	25.7482	M	3-7
40	11.7915	M	0-4	85	17.8014	M	13-1	130	22.2178	M	5-5	175	25.8260	E	2-8
41	12.2251	M	8-1	86	17.9598	M	2-5	131	22.2191	E	20-1	176	25.8912	E	9-5
42	12.3386	M	5-2	87	18.0155	E	1-6	132	22.4010	E	4-6	{177	25.9037	M	1-8
43	12.6819	E	4-3	88	18.0633	E	16-1	133	22.5014	E	9-4	{178	25.9037	E	0-8
44	12.8265	E	11-1	89	18.0711	M	0-6	134	22.5827	M	3-6	179	26.1778	E	15-3
45	12.9324	E	7-2	90	18.2876	M	7-3	135	22.6293	E	12-3	180	26.2460	E	12-4

* Nomenclature after Barrow & Mieher, "Natural Oscillations of Electrical Cavity Resonators," IRE Proceedings, April 1940, p. 184. M modes take zeros of $J_l(x)$; E modes take zeros of $J'_l(x)$. Number directly following E or M is l; number after hyphen is number of root.

Values less than 16.0 are abridged from six-place values and are believed to be correct; value more than 16.0 are abridged from five-place values and may be in error by one unit in fourth decima place. **5** in fourth place indicates that higher value is to be used in rounding off to fewer decimals

One solution, of limited application, is to choose an operating rectangle free from extraneous responses. An alternative solution is to design the cavity in a manner such that the undesired responses are reduced or sup-

pressed to a point where their presence does not interfere with the normal operation of the cavity. In this latter case, the amount of suppression is naturally dependent upon the use to which the cavity is to be put, and is conceivably different for a high Q cavity used as a frequency meter, for example, and one used as a selective filter.

Experience has shown that certain families of modes are much more difficult to suppress than others, and are to be avoided, if at all possible. The feasibility of doing this can be determined by sliding the operating area (a suitable opening in a sheet of paper) around on a large mode chart until the most favorable operating region, consistent with other requirements, has been found.

Once a suitable operating area has been chosen, the cavity diameter is fixed and length and frequency scales added to the mode chart make it read directly in quantities readily measured.

Cavity Couplings

To be useful the cavity must be coupled to external circuits. The problem here is to get the correct coupling to the main mode and as little coupling as possible to all others. Since the electric field is zero everywhere at the boundary surface of the cavity for the $TE\ 01n$ mode, coupling to it must be magnetic. This may be obtained either by a loop at the end of a coaxial line or by an orifice connecting the cavity with a wave guide.

The location for maximum coupling to the main mode is on the side of the cavity, an odd number of quarter-guide wavelengths from the end, or on the end about halfway (48%) out from the center to the edge. Correct orientation of loop or wave guide is achieved when the magnetic fields are parallel. This requires the axis of the loop to be parallel to the axis of the cylinder for side wall feed and to be perpendicular to the cylinder axis for end feed. Wave guide orientation is shown in Table IV.

The theory of coupling loops and orifices is not at present precise enough to yield more than approximate dimensions. Exact sizes of loops and holes have therefore been obtained experimentally for all designs.

On the basis of rather severely limiting assumptions,[11] coupling formulas for a round hole connecting a rectangular wave guide and a $TE\ 01n$ cavity are given in Table IV. The assumptions are that the orifice is in a wall of negligible thickness, its diameter is small compared to the wavelength, it is not near any surface discontinuity, and that the wave guide propagates only its principal (gravest) mode and is perfectly terminated. In echo box applications, this theory leads to a computed diameter that is somewhat smaller than experiment shows to be correct.

The coupling to other modes can be analyzed, at least qualitatively, from the field expressions of Table I. This has been of value in making final

	CASE 1A	CASE 1B	CASE 2A	CASE 2B	CASE 3A	CASE 3B
COUPLING METHOD						
CIRCULAR ORIFICE	$\dfrac{\Delta f}{f} = -K_C \dfrac{\lambda^2 d^3}{D^4 L}$ $W_a = K_W \dfrac{\lambda^2 d^6}{\lambda_g w h D^4 L}$	$\dfrac{\Delta f}{f} = -K_C \dfrac{n^2 \lambda^2 d^3}{D^2 L^3}$ $W_a = K_W \dfrac{n^2 \lambda^2 d^6}{\lambda_g w h D^2 L^3}$	$\dfrac{\Delta f}{f}$ SAME AS 1A $W_a = K_W \dfrac{\lambda_g \lambda^2 d^6}{w^3 h D^4 L}$	$\dfrac{\Delta f}{f}$ SAME AS 1B $W_a = K_W \dfrac{n^2 \lambda_g \lambda^2 d^6}{w^3 h D^2 L^3}$	$\dfrac{\Delta f}{f}$ SAME AS 1A W_a SAME AS 1A	$\dfrac{\Delta f}{f}$ SAME AS 1B W_a SAME AS 1B
CONSTANTS	K_C K_W TE 01n 0.316 1.322 TE 02n 1.058 4.43 TE 03n 2.225 9.32	K_C K_W K_m 0.1107 0.464 0.2403 0.1995 0.836 0.1312 0.288 1.207 0.0905	K_W 0.331 1.108 2.330	K_W 0.1159 0.2089 0.302		

NOTATION: λ = FREE SPACE WAVELENGTH d = DIAMETER OF ORIFICE
OF CAVITY RESONANCE $W_a = \dfrac{1}{Q_a}$ = CAVITY LOADING
λ_g = GUIDE WAVELENGTH

NOTE: FOR FEED LIKE CASES 2′ AND 3, BUT WITH WAVEGUIDE TERMINATED IN BOTH DIRECTIONS. DIVIDE W_a BY 2

TABLE IV.—*Orifice Coupling of Wave Guide (TE 10 Mode) to Cylindrical Cavity (TE 0mn Mode)*

small relocations of the coupling points to discriminate against a residual undesired mode.

Principle of Similitude

One other theorem generally applicable to all cavities has been useful in design. It is the principle of similitude, which may be stated as follows[12]:

A reduction in all the linear dimensions of a cavity resonator by a factor $1/m$ (if accompanied by an increase in the conductivity of the walls by a factor m) will reduce the wavelengths of the modes by a factor $1/m$.

The condition given in parentheses is necessary for strict validity; for high Q cavities, it need not be considered.

APPLICATION OF THEORY

An illustration of an engineering application of the basic information just presented, is the design of an echo box test set for use in radar maintenance.

The test set has a number of components, but only the cavity proper and its couplings will be considered at this time.

Design Requirements

The basic design requirements of the cavity set by the radar are: (1) the working decrement, and (2) the tunable frequency limits (f_1 and f_2). Service use of these test sets in the 3 *kmc* and 9 *kmc* bands has shown that a working decrement of about 3 *db* per microsecond gives satisfactory results.

As seen in the discussion above, Q is a more useful design parameter for the resonant cavity than decrement, d. Hence the conversion

$$Q = \frac{27.3f}{d} \tag{4}$$

where d is expressed in *db* per microsecond, and f is in megacycles, gives the loaded or working Q to be realized.

Determining the Theoretical Q Required

For design purposes, however, there are several factors which dictate the use of a value for theoretical Q which is somewhat higher than the loaded Q just computed. The input and output couplings reduce the theoretical Q of the cavity due to their loadings. The coupling factor, s, expressed as the ratio of loaded to unloaded Q, has values for echo boxes of about 0.90 for the input coupling and from 0.90 to 0.99 for the output coupling. In addition, other factors such as the means used for mode suppression may degrade

the Q. But, in simple designs, these may be negligible. Therefore, it is expedient to design for a theoretical Q of about 15 to 25 per cent in excess of the working Q to be realized.

The working Q's of a number of echo box designs are cited here to indicate the order of magnitude required at several frequencies:

1 kmc	70,000*
3	40,000
9	100,000
25	200,000

Finding the Cavity Dimensions

With the frequency and theoretical Q known, the dimensions of the cavity can be evaluated but the formulas of Table I require some simplification for engineering use.

The mode-shape (MS) factor, $Q\frac{\delta}{\lambda}$, may also be termed its selectivity which for the $TE\,01n$ modes may be expressed as follows:

$$Q\,\frac{\delta}{\lambda} = 0.610\,\frac{\left[1 + 0.168\left(\frac{D}{L}\right)^2 n^2\right]^{3/2}}{1 + 0.168\left(\frac{D}{L}\right)^3 n^2} \tag{5}$$

where:

δ = the skin depth in cm. = $\dfrac{1}{2\pi}\sqrt{\dfrac{10^3\rho}{f}}$

ρ = the resistivity in ohm-cm.

The skin depth is a factor which recognizes the dissipation of energy in the walls and ends of the cylinder. With increase of resistivity of these surfaces the currents penetrate deeper and the resulting Q is lower.

A comparison of the relative Q's computed from the resistivity of several metals will show the importance of this factor:

Silver	1.03
Copper	1.00
Gold	0.84
Aluminum	0.78
Brass	0.48

Therefore, a brass cavity will have about one-half of the Q that a similar cavity would have if made of copper. Similarly, the silverplating of a copper cavity will gain about 3 per cent in Q.

Equation 5 may be made more convenient for calculations by combining

* This value reflects the higher Q required on ground radars.

terms which are a function of frequency and by assuming the conductivity of copper* for the cylinder walls. It then becomes

$$\frac{Q\sqrt{f}}{10^6} = 2.77 \frac{\left[1 + 0.168\left(\frac{D}{L}\right)^2 n^2\right]^{3/2}}{1 + 0.168\left(\frac{D}{L}\right)^3 n^2} \tag{6}$$

for $TE\,01n$ modes; f is in megacycles.

Thus, it is seen that for the design parameters Q and f, selected values of n now define tentative useful points on the mode chart in terms of D/L and n.

Selection of Operating Area on Mode Chart

This will be more evident upon examination of Fig. 6, which is a basic design chart for cylindrical cavity resonators using $TE\,01n$ modes. The coordinates of $(fD)^2$ and $(D/L)^2$ will be recognized from the previous discussion of the mode chart (Fig. 1) although in this case the range has been expanded by the use of logarithmic coordinates. Mode identification is obtained from equation 2; which, for $TE\,01n$ modes becomes

$$(fD)^2 \times 10^{-8} = 2.0707 + 0.3480\, n^2\, (D/L)^2 \tag{7}$$

with D and L in inches and f in megacycles.

A family of $TE\,01n$ modes has been drawn on the chart for selected values of n. To aid in designing for minimum volume a line labelled Max. $\frac{Q}{V}$ has been added (Equation 3). Lines of constant $Q\sqrt{f}$ are also shown as a series of dashed lines.

A tentative operating area on this chart may be selected on the basis of the required $Q\sqrt{f}$. Using mid-frequency for f, the intersection of the $Q\sqrt{f}$ line and the minimum volume line will define the operating mode, n, and also locate the center of the operating rectangle.

An enlarged plot of this area as in Fig. 7 will show all modes possible in the cavity. Some adjustment of the precise location of this area may then be desirable to eliminate certain types or numbers of unwanted modes. For example, it is extremely difficult to suppress $TE\,02$ modes without affecting the $TE\,01$ mode, because of the very close resemblance of the field configurations. On the other hand, most TM modes are easy to handle.

It may be possible to select an operating area such as the rectangle blocked out in Fig. 1 in which all extraneous responses (with the exception of the companion $TM\,111$) are avoided. The largest rectangle which can be inscribed here is limited by the $TE\,311$ and $TE\,112$ modes. This will

* $\rho = 1.7241 \times 10^{-6}$ ohm-cm—the International Standard value for copper.

permit a ± 3.6 per cent frequency range. Several designs of 3 kmc echo boxes have been based on this area.
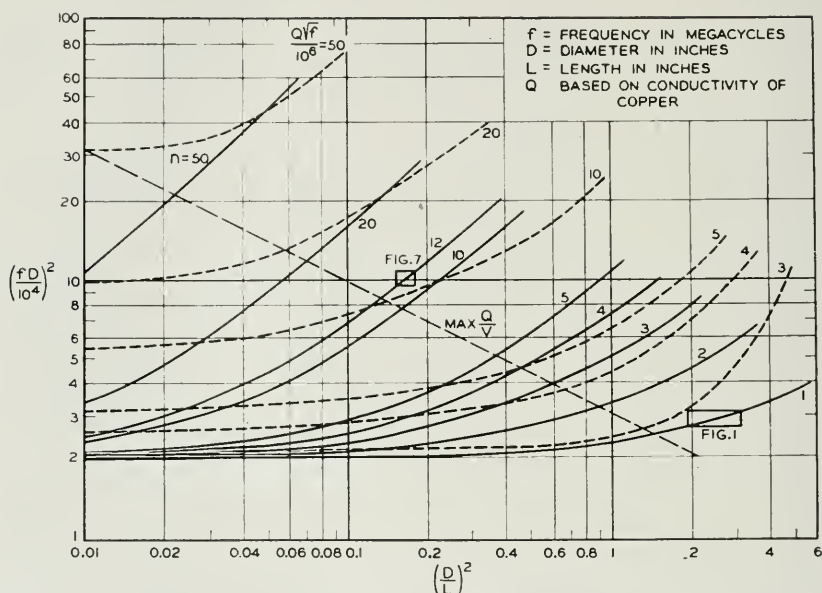

Fig. 6—Design chart for cylindrical cavity resonators operating in the TE $01n$ modes.
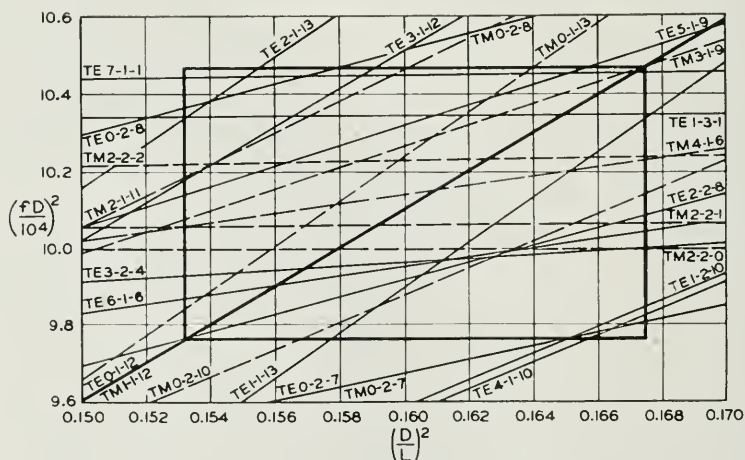
Fig. 7—Expanded mode chart covering the operating area of a 9 kmc echo box.

This type of solution is quite adequate for many of the simpler designs; but as the required Q becomes larger or the frequency coverage becomes

greater it is generally not possible to locate such areas. An operating area for 9 *kmc* is shown on Fig. 7. Nine crossing modes and twelve interfering modes exist. For 25 *kmc* the crossing modes run into the forties with hundreds of interfering modes.

Suppression of the undesired modes requires a thorough knowledge of their field configurations and a number of effective techniques which may be applied on a practical engineering basis. Several examples are cited in later sections.

Since decrement is an important characteristic of these cavities, especially when applied to radar test sets, the uniformity of the decrement over the frequency range or "flatness of response" may be a significant design requirement. It will be seen from Fig. 3, that the MS factor of the wanted mode is not constant with varying D/L. In fact, if it were, the decrement would increase as the 3/2 power of frequency.

There are at least three attacks on this "flatness" problem: (1) to operate on the sloped portion of the MS curve in such a manner that its characteristic will tend to be complementary to the change with frequency; (2) to obtain compensation by varying the coupling with frequency—generally accomplished by selecting an appropriate coupling point along the side wall of the cavity; and (3) to overplate a portion of the cylinder's interior surface with a material of higher resistivity such as cadmium. For this third method, the formulas for $Q\frac{\delta}{\lambda}$ of Table I are no longer applicable since they assume a uniform resistivity of the cavity walls.

Thus, it will be seen that the final design of a cavity resonator is a compromise between a number of desired characteristics:

a) A cavity of minimum volume for a given Q.

b) A cavity having a minimum of extraneous responses of types difficult to suppress.

c) A cavity with compensation for flatness of decrement.

Engineering judgment is required to weigh the emphasis on each of these requirements which at times may be mutually exclusive.

Some Practical Considerations

Physically realizing the theoretical characteristics just described to obtain a satisfactory cavity brings forth a host of practical design problems. A number of these will be discussed in this section.

Description of Echo Box Test Sets

The schematics (Figs. 8 and 9) and photographs (Figs. 12 to 14) show the components and various construction methods of echo boxes in the 3 and 9 *kmc* bands. The cavity itself may be spun, drawn or turned of material

such as brass or aluminum which will give mechanical stability without excessive weight.
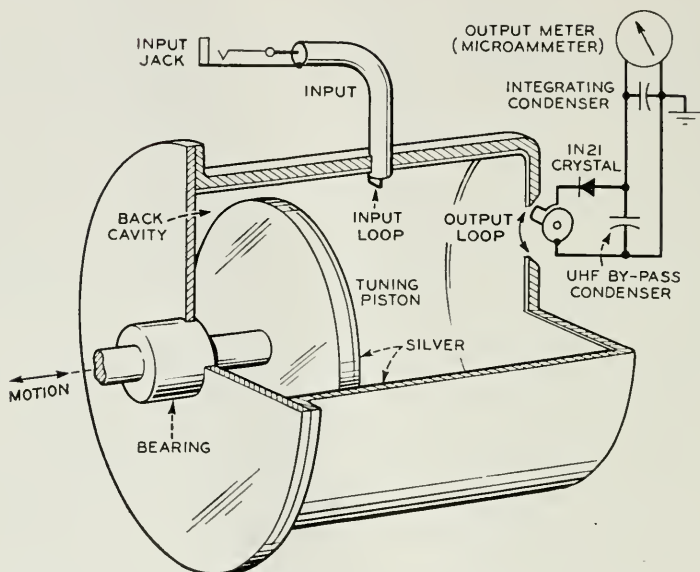
Fig. 8—Schematic of a 3 *kmc* echo box.
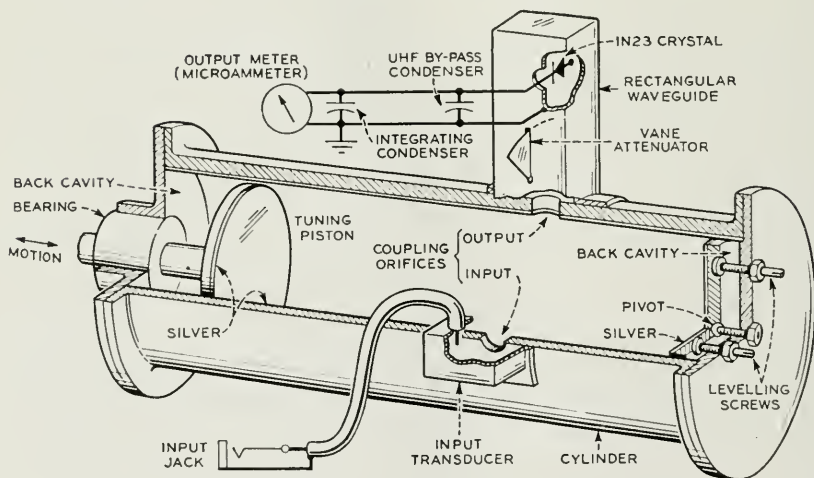
Fig. 9—Schematic of a 9 *kmc* echo box.

The movable plate is driven by the piston which operates in a close-fitting bearing. The drive mechanism translates the rotary motion of the tuning

knob into linear motion of the piston through a crank and connecting rod assembly. Coupled to the drive shaft is an indicating dial to register frequency.

Silverplating is indicated on all interior surfaces of the cavity for minimum resistance to currents in the walls and ends of the cylinder.

For the 3 *kmc* bands, the input coupling is in the form of a loop protruding into the cavity and connected by coaxial microwave cable to the radar under
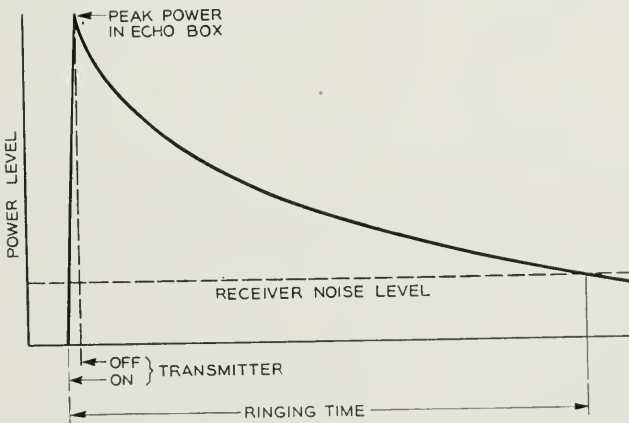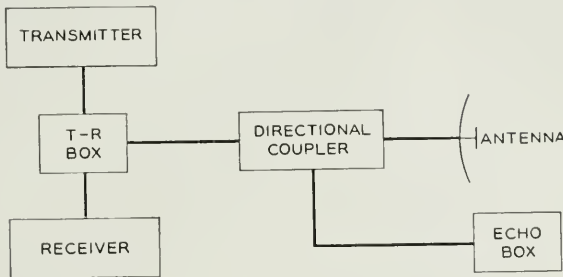
Fig. 10—Radar test with echo box.

test. For the 9 *kmc* bands, the input coupling is an orifice with which is associated a wave guide to coaxial cable transducer.

The output couplings for the two frequency bands also differ in construction. For the 3 *kmc* bands, the variable coupling is achieved by rotating the output loop before an aperture in the cavity. In this rotary mount is housed the rectifying crystal and by-pass condenser. An orifice is used for the 9 *kmc* band output coupling which feeds a crystal mounted in a wave guide. Amplitude control is by a vane attenuator in the wave guide. In

all cases an integrating condenser is required to smooth out the *d-c* pulses delivered by the crystal to give output indication on the *d-c* microammeter.
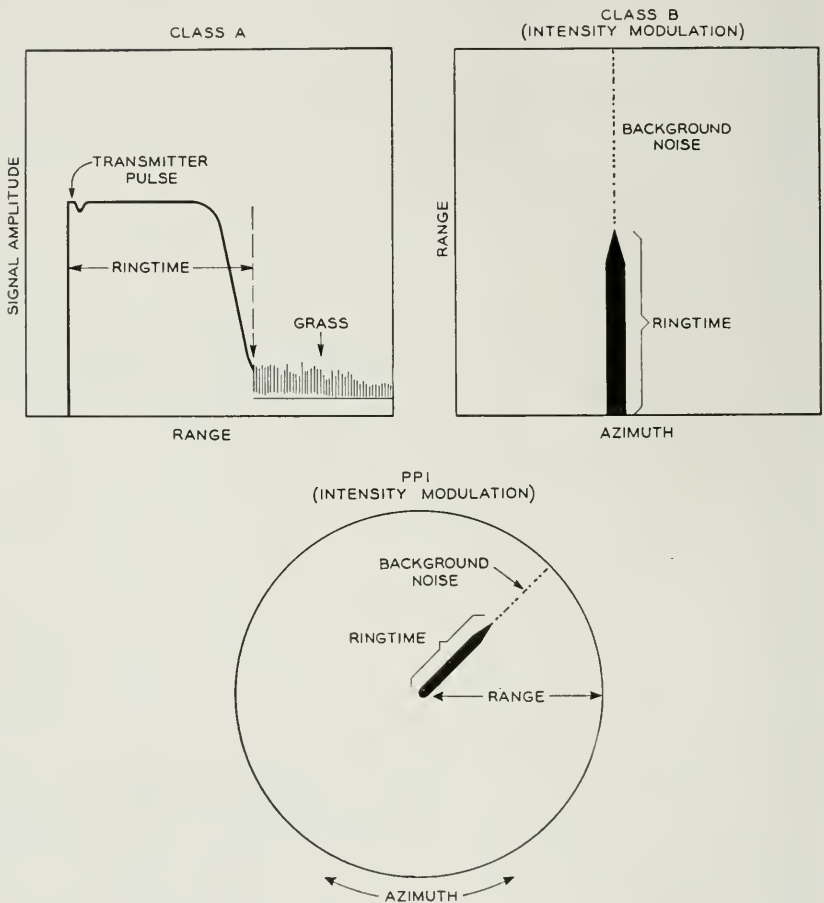
Fig. 11.—Typical ringtime patterns on radar indicators.

The cylinder in the 9 *kmc* band is an assembly of a tube and two end plates. One of these is driven by the piston as described and the other is adjustable for "levelling" i.e., parallelism of the two plates.

End Plate Gaps

The wanted *TE* mode is paired with a companion unwanted *TM* as described above. The fact that the *TM* mode has a *Q* substantially less than that of the *TE* makes the realization of the higher *Q* difficult. A method is

needed of suppressing the *TM* in the presence of the *TE*. Since the resonant cavity is a cylinder, of variable length, the movable end plate (or reflector)
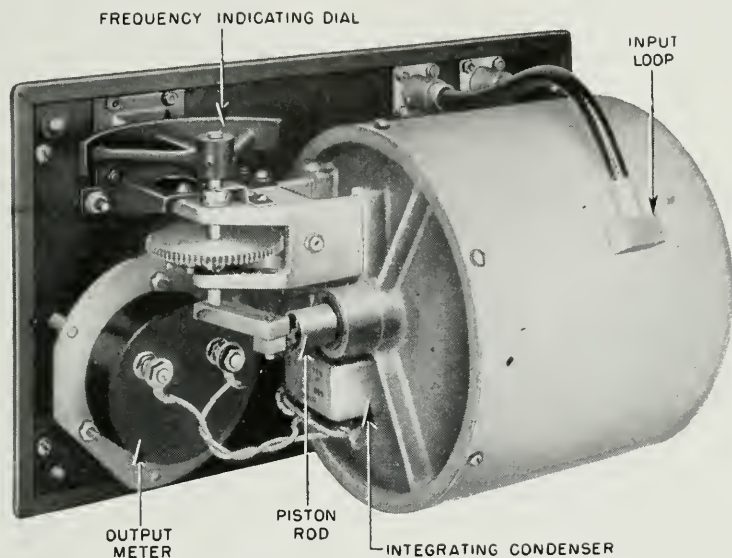
Fig. 12—Type of construction in a *3 kmc* echo box.
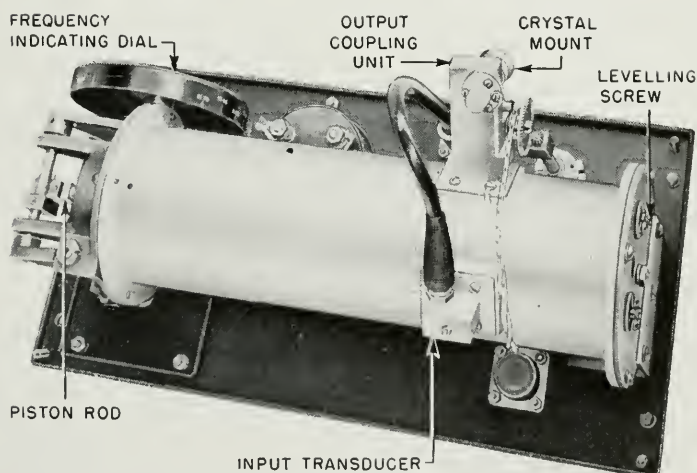
Fig. 13—Type of construction in a *9 kmc* echo box.

can be modified to include a gap at its periphery. The gap perturbs the resonant frequencies of the two modes by different amounts so that they

become separated.   Secondly, the peripheral gap cuts through the surface
currents at points of high density for TM modes and minimum density of
$TE\ 0mn$ modes and hence is a form of mode suppression.   Thirdly, the gap
greatly simplifies the mechanical design of a movable end plate by elimi-
nating the need of physical contact with the side wall of the cylinder.

A similar gap may be used at the other end.   This facilitates "levelling"
of a false bottom in the cavity.
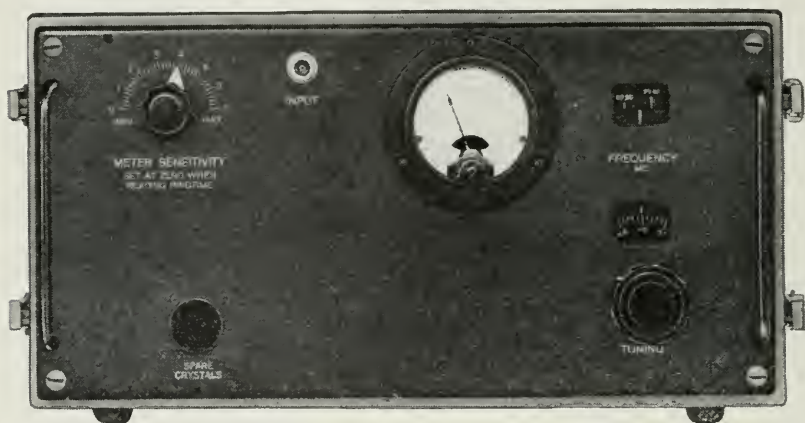
Fig. 14—Panel view of a 9 kmc echo box.

Back Cavity Effects

The cavity with a peripheral gap may give rise to further spurious reso-
nances in the region behind the reflecting surface (known as the back cavity)
if these responses are not damped.   The addition of a lossy material such
as bakelite or carbon loaded neoprene in the back cavity is a successful sup-
pression method.

Cylinder Tolerances

The geometry of the structure is very important in realizing the poten-
tial Q of the cavity.   The theoretical computations are based on a perfect
right circular cylinder which in practice is seldom achieved.   Distortions
occur in various forms: e.g., the cylinder instead of being round may be
elliptical; the ends may not be perpendicular to the axis of the cylinder or
not parallel to each other; surface irregularities may be present causing field
distortions within the cavity.   Many of these effects have been minimized
by requiring adherence to close dimensional tolerances.

For some designs, the requirement on the parallelism of the end plates is
greater than can be commercially produced.   An adjustable mechanism for

levelling is a practical solution. Tilt adjustments in the order of 0.001 inch at the edge of the plate (about 3-inh diameter) are required in the 9 *kmc* band. Experience in the 25 *kmc* band shows evidence of the need for even finer control.

Plating

In addition, adequate control of the conductivity of the interior surfaces of the cavity is necessary to achieve a uniform manufactured product. This requires attention not only to the thickness and uniformity of the plating but also to the purity of the plating baths and the avoidance of introduction of foreign matter during buffing processes.

Couplings

The type and location of the coupling means can be used to discriminate between wanted and unwanted modes. Hence, this is a fertile field for mode suppression techniques. For example, since TM modes have $H_z = 0$, orifice coupling to the main mode at the side wall of the types shown in Table IV, cases 1A and 2A, will not couple to any TM modes. Again, if end coupling is used in a cavity which will support both the TE 01 and TE 02 modes, by locating this coupling at the point where $H_\rho = 0$ for the TE 02 mode (about 54% of the way out from the center), it will not be excited and coupling to the TE 01 will be only slightly below maximum.

For echo box test sets the magnitude of the input coupling to the wanted mode is a compromise between the incomplete buildup of the fields within the cavity during the charging interval and the loading of the cavity Q on discharge. This is carried on by varying the coupling and observing the "ringtime" (the echo box indication on the radar scope). Optimum coupling is achieved when ringtime is made a maximum.

Output couplings for echo boxes are made so that just enough energy is withdrawn from the cavity to give an adequate meter reading.

Drive Mechanism

The objective of the design of the tuning mechanism is to provide a smooth, fine control with a minimum of backlash. An illustration of the mechanical perfection required can be cited in a 9 *kmc* band design where $\frac{1}{4}$ inch of travel covered 200 megacycles in frequency. Hence, for frequency settings to be reproducible to within $\frac{1}{4}$ mc the mechanical backlash of the moving parts had to be held to about 0.0003 inches or 0.3 mil. To realize this in commercial manufacture and to maintain it after adverse operating conditions such as vibration and shock was a major mechanical design problem.

In the design of this drive mechanism it should be recognized that equal

increments of cavity length will not produce uniform increments in frequency. The mode chart indicates graphically that a straight-line relationship exists between $(f)^2$ and $\left(\dfrac{1}{L}\right)^2$. Uniformly spaced markings on a dial reading directly in frequency can be realized by the use of such mechanisms as an eccentric operating on a limited arc. Adjustments are customarily provided to bring the cavity resonance and dial indication into agreement at some frequency of test. Frequency departures of the drive mechanism referred to this point are held commercially to about one part in 5000.

Application of Similitude

Echo box developments have often been undertaken at frequencies where adequate test equipment was not available. This has been especially true as the radar art progressed to higher and higher frequencies.

The principle of similitude has been utilized in the construction and test of models at the frequency of existing test facilities. The models have then been scaled to the assigned frequency band. This has been found to be a very practical expedient.

Use of Cavities For Radar Testing

The high Q resonant cavity when appropriately connected to a radar system returns to it a signal which may be used to judge the over-all performance of the radar. Its operation is as follows: During the transmitted pulse, microwave energy from the radar is stored in the cavity in the electromagnetic field. The charge of the cavity increases exponentially during this interval but fails to reach saturation for the cavity by a substantial margin because the pulse is too short. At the end of the pulse, the decay of this field supplies a signal of the same frequency as that of the radar transmitter (when the echo box is in tune) which is returned to the radar receiver as a continuous signal diminishing exponentially in amplitude.

The time interval between the end of the transmitted pulse and the point where the signal on the radar disappears into the background noise is the "ringtime." The term is used somewhat loosely since, in actual practice, the ringtime is measured on indicators whose range markings are generally in miles or yards referred to the beginning rather than the end of the pulse. The difference, of course, is small. It is customary to include the pulse length in all ringtime figures on operating radar systems. Typical ringtime patterns on radar indicators and a schematic of a radar test with an echo box are shown in Figs. 10 and 11.

As the output power of the radar either increases or decreases corresponding changes in the "charge" of the cavity will be reflected directly in ringtime changes. Similarly as the noise level of the radar receiver varies the

merging point of the cavity signal and the noise will show proportional changes in ringtime. Hence, the ringtime indication measures these two factors on which the radar's ability to discern real targets so largely depends.

The exponential buildup and decay of the charge in the cavity occur at a rate determined by the working decrement of the cavity. As mentioned previously, a decrement of about 3 *db* per microsecond is a satisfactory value for the 3 *kmc* and 9 *kmc* bands. A one microsecond change in ringtime (roughly one-tenth mile) would, therefore, represent a change in system performance of 3 *db*.

Uniformity Control and Expected Ringtime

By introducing an adjustment for the working Q of the cavities it is now possible to control the uniformity of the manufactured product to very close limits. Other improvements have also been incorporated which insure that boxes which have been made alike as to Q will similarly give uniform ringtime indications on a test radar. If the test sets are all alike as to ringtime, it is then possible to quote an "expected ringtime" for each of the various radars to be serviced by the echo box. Initially a measuring tool indicating relative changes in day to day operation of the radars, the uniformity provision with its "expected ringtime" has made the echo box test set an absolute measuring instrument of moderate precision.

Other Uses

In addition to its use as a measure of over-all performance of a radar, a significant number of diagnostic tests may be performed when trouble develops, which aid in rapidly locating the source. One such test is spectrum analysis. The extreme selectivity of the high Q cavity permits examination of the spectrum of the pulsed wave and from this may be deduced characteristics of the pulse, including pulse length. Multiple-moding of the magnetron circuit is easily shown by this analysis.

The meter of the test set gives a relative indication of the output power of the radar and this in itself assists greatly in segregating transmitter troubles from receiver troubles.

Also of importance is the use of the echo box as a frequency meter. The high Q of the cavity plus the fine control of the drive mechanism and the direct reading dial give excellent results (comparable to that of a wavemeter).

Acknowledgement

The design of resonant cavities is a difficult and complex art. In bringing it to the present state, the number of individuals who have made significant mathematical, theoretical, engineering and mechanical contributions is so

large that it is impossible to give them due credit by name. To them, however, the authors wish to acknowledge their indebtedness and to express their appreciation.

BIBLIOGRAPHY

1. E. I. Green, H. J. Fisher and J. G. Ferguson, "Techniques and Facilities for Microwave Radar Testing", A. I. E. E. Transactions, 65, pp. 274–290 (1946), and this issue of the *B. S. T. J.*
2. G. C. Southworth, "Hyper-Frequency Wave Guides—General Considerations and Experimental Results," B. S. T. J., 15, pp. 284–309 (1936).
3. J. J. Thompson, "Notes on Recent Researches in Electricity and Magnetism," Oxford, 1893.
4. Lord Rayleigh, "On the passage of electric waves thru tubes or the vibration of dielectric cylinders," *Phil. Mag.*, 43, pp. 125–132 (1897).
5. A. Becker, "Interferenzrohren für electrische Wellen," *Ann. d. Phys.*, 8, pp. 22–62 (1902).
6. F. K. Richtmyer and E. H. Kennard, "Introduction to Modern Physics," pp. 184–189, McGraw Hill, 1942.
7. R. I. Sarbacher and W. A. Edson, "Hyper and Ultra-high Frequency Engineering," p. 377, Wiley, 1943.
8. H. A. Wheeler, "Formulas for the Skin Effect," *Proc. I. R. E.*, 30, pp. 412–424 (1942).
9. W. L. Barrow and W. W. Mieher, "Natural Oscillations of Electrical Cavity Resonators," *Proc. I. R. E.*, 28, pp. 184–191 (1940).
10. L. J. Chu, "Electromagnetic Waves in Elliptical Hollow Pipes of Metal," *Jour. App. Phys.*, 9, pp. 583–591 (1938).
11. H. A. Bethe, M.I.T. Radiation Laboratory Reports V-15-S, 43–22, 43–26, 43–27, and 43–30.
12. H. Konig, "The Laws of Similitude of the Electromagnetic Field, and Their Application to Cavity Resonators," *Hochf:tech u. Elek:akus*, 58, pp. 174–180 (1941). Also *Wireless Engr.*, 19, pp. 216–217, No. 1304 (1942).

Techniques and Facilities for Microwave Radar Testing*

By E. I. GREEN, H. J. FISHER, J. G. FERGUSON

Methods and devices are described for testing microwave radars in the radio frequency range from about 500 mc to 25,000 mc, and at associated video frequencies. In general, the same instruments and techniques are applicable also in testing microwave communication systems.

INTRODUCTION

THAT radars are marvels of ingenuity has long since become common knowledge. This ingenuity is reflected, however, in complexity of circuits. A rough index of this is found in the number of vacuum tubes, which for a single radar may range from 50 to 250. Notwithstanding the most careful design, it is easy for the radar performance to become impaired under operating conditions.

Not only is radar complex, but its performance criteria are less tangible than those of conventional communication systems. Ordinary radio is to some extent self-testing in that reception of intelligible speech or signals frequently constitutes a sufficient check of satisfactory performance. With radar, the greater the range coverage and the more accurate the data, the more valuable the information is likely to be. However, the working range may fall to a fraction of the possible maximum or some other degradation or malfunctioning may occur, with nothing in the operation of the radar to tell that this has happened. Since lack of maximum performance may have serious military results, measurement of performance assumes the utmost importance in radar work.

The new techniques and new frequency ranges employed for radar necessitated the wartime development of a wide variety of new types of test equipment. A large part of this development work was concentrated at Bell Laboratories and at the N.D.R.C.'s Radiation Laboratory at M.I.T., working in close coordination with one another and with the technical services of the Army and Navy. In the manufacture of radar test equipment, Western Electric took a major part. This article discusses the techniques of radar testing and describes the types of test gear developed by Bell Laboratories and manufactured by Western Electric. These cover the radio frequency range from about 500 megacycles to 25,000 megacycles, together with associated video frequencies.

Because of its importance during the war, emphasis has been placed on

the testing of microwave radars. However, similar methods and instruments have also been employed in the testing of microwave communication systems and such applications can be expected to increase. In this situation the developers and users of microwave communication systems are fortunate in that almost all of the techniques and devices developed for radar testing are equally applicable to communication systems. This is even true of the video units, which are useful in connection with pulse modulated telephone systems and AM or FM television systems.

So many persons, both within and outside Bell Laboratories, have contributed to the developments described that the authors have reluctantly reached the conclusion that the assignment of individual credit should not be attempted.

Requirements

Operation of Typical Radar

Subsequent discussion may be simplified by first reviewing briefly the operation of a somewhat typical radar, as shown in Fig. 1. Under the control of d.c. or so-called video pulses from the modulator, short pulses of radio frequency energy are delivered by the magnetron transmitter to a highly directive antenna, ordinarily arranged to scan a section of space. Energy reflected from an object or "target" in the path of the beam is intercepted by the same antenna. The received pulses or echoes are converted to an intermediate frequency by heterodyning against a local oscillator, the frequency of which may be automatically controlled.

To enable the same antenna to serve for both transmitting and receiving, a TR tube or transmit-receive switch is usually provided. This consists of a partially evacuated resonant cavity containing a spark gap which breaks down during the transmitted pulse, thus preventing the transmitted power from injuring the sensitive receiver. An RT tube, consisting of a similar resonant cavity and spark gap, may be provided to prevent absorption of the received signal by the transmitter. After amplification and detection of the received signal, the resultant video pulses are applied to an indicator which may present information in any of several different ways. Customarily the direction of the target (determined by antenna orientation) and its range (determined by reflection interval, 10.7 microseconds per mile) are shown. The system may be used merely for searching, or for fire control, bomb direction, or other functions, with additional equipment as required.

Types of Tests and Test Sets

Figure 2 shows an early assemblage of radar test equipment for the 10 cm range, initially produced in 1942, which has seen wide usage.

The more important types of tests required in radar work, either at radar operating locations or at centralized service points, are: (1) Over-all Performance (Range Capability), (2) Transmitter Power, (3) Receiver Sensitivity, (4) Transmitter Frequency, (5) Transmitter Spectrum, (6) Standing Wave Ratio, (7) R. F. Envelope, (8) Receiver Recovery, (9) AFC Tracking, (10) I.F. Alignment, (11) Video Wave Shapes, (12) Range Calibration and (13) Computer Calibration.

The principal types of test sets required for carrying out the above are: (1) Signal Generators, (2) Echo Boxes, (3) Frequency and Power Meters (separate or combined), (4) Standing Wave Meters, (5) R.F. Loads, (6)
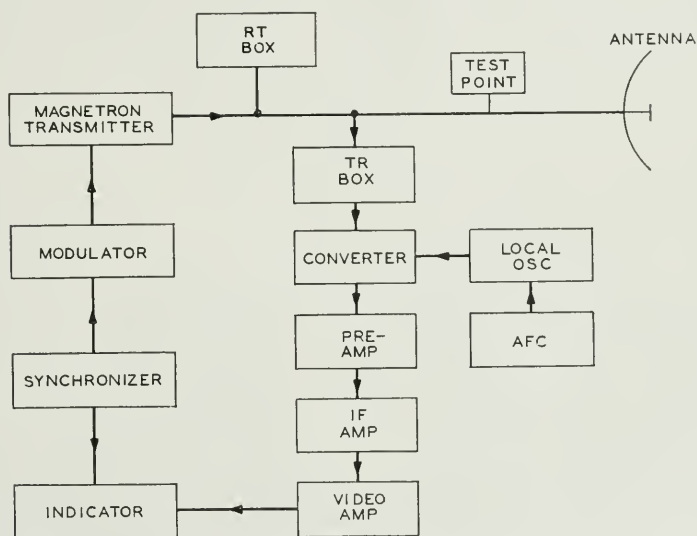
Fig. 1—Diagram of typical radar.

Oscilloscopes, (7) Video Dividers and Loads, (8) Range Calibrators, (9) Computer Test Sets and (10) Spectrum Analyzers. Various auxiliary testing instruments are also employed, including vacuum tube testers, I.F. signal generators, audio oscillators, flux meters, etc. Before discussing the above tests and devices individually, some of the requirements for radar test equipment, especially those resulting from military usage, will be summarized.

Generality of Application

Radars perform a great variety of functions, including search and surveillance, gun laying and fire control, bomb direction, and navigation. They are used in the air, on shipboard, and on the ground, in attack and in de-

fense, in combat zones and in rear areas. To realize the advantages of different parts of the frequency spectrum, avoid interference, and keep ahead of enemy countermeasures, it has been necessary for radar to exploit many different frequency bands and sub-bands. These diversities have
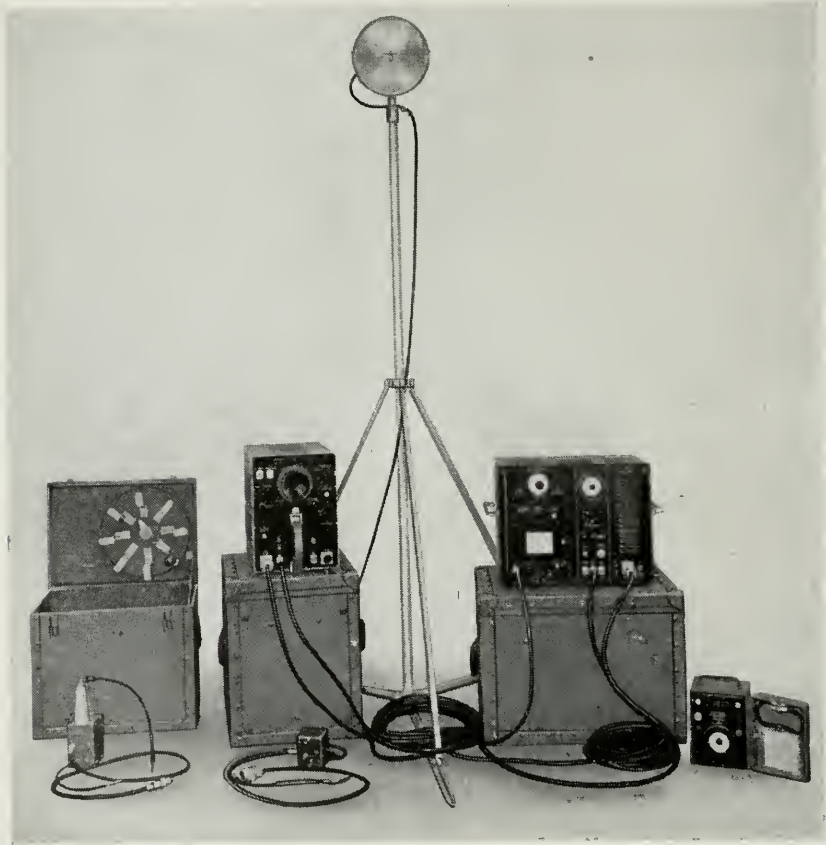
Fig. 2—Early radar test set for 3000 mc range—includes signal generator, oscilloscope, power meter, test antenna and auxiliary units.

bred a multiplicity of types of radar, with a corresponding variety of testing problems and requirements.

Maintenance and testing of radars must be performed at many different locations. In the Army these locations, known as echelons, include (a) the operating unit, (b) central service points, either fixed or mobile, for a number of operating units, and (c) large depots either in the military theatres or in the United States. Navy maintenance and testing are carried out on board the fighting ships or auxiliaries, and at advance and overhaul bases.

In developing test equipment, an offhand approach might have been to provide specialized equipment for testing each radar under specific conditions. Since this would have made the total burden of development, manufacture and field maintenance well nigh intolerable, a coordinated plan was followed whereby (with minor exceptions) each test set was made capable of widespread application in testing as many radars under as varied conditions as possible.

Broadbanding

Generality of application required the designing of test equipment for broad frequency bands, bands as a rule much wider than those of the radars themselves. It was necessary, therefore, not only to develop new microwave testing techniques, but to advance the art still further to render the testing components as far as possible insensitive to frequency.

Precision

A radar itself is an instrument of considerable precision. The test equipment used for checking the radar performance in the field has to have still higher accuracy. It is noteworthy that the measuring accuracy realized throughout the microwave range is comparable with that obtainable at lower frequencies where many years of background exist.

Packaging—Size and Weight

Light weight and compactness are of paramount importance where a test set has to be carried any distance by the maintenance man, where it is used in cramped quarters in a plane, truck or submarine, or where it has to be taken up ship ladders or through small hatchways. To permit portable use under such conditions, the design objective was established of a weight not exceeding about 30 lbs. (exclusive of transit case), combined with a ruggedness adequate for all conditions of use. Through rigorous attention to both mechanical and electrical design, this objective has been realized (in many cases with considerable margin) except for a few sets intended primarily for bench use. Figure 3 illustrates the use of lightweight test equipment in maintaining airborne radars.

Environmental Influences

Military usage requires that the test equipment be capable of efficient operation at any ambient temperature between a minimum of the order of $-40°$ to $-55°$ C and a maximum of the order of $+65°$ to $+70°$ C, as well as at any relative humidity up to 95%. In addition the set must withstand continued exposure to driving rain, dust storms and all other conditions encountered in tropical, desert or arctic climates. Often the test set in its

transit case must be capable of submergence under water without ill effects. Fungus-proofing with a fungicidal lacquer is a standard requirement.

Simplicity, Reliability, Accessibility

Not only must the functioning of the test set be reliable, stable and trouble-free, but the set must make minimum demands for special skill or tech-

Fig. 3—Portable units used in checking airborne radars at Boca Raton, Fla.

nique on the part of the maintenance man. Access for maintenance purposes, while important in radar test sets, is not as controlling as in the radars themselves and sometimes has to be sacrificed in part for compactness.

Ruggedness

For general application, the test equipment must be capable of withstanding airplane vibration, the shock of heavy guns, depth charges and

near misses, and the combinations of shock and vibration connoted by the requirement of "transportation over all types of terrain in any Army vehicle." Test and experience have made it possible to translate these general requirements into two specific requirements, namely the ability to withstand (1) vibration at frequencies from 10 to 33 cycles per second with $\frac{1}{16}''$ excursion for 30 minutes in each of three axes and (2) the shock produced by a 400 lb. hammer falling through distances of 1, 2 and 3 ft. in each of three axes, and striking an anvil to which the set is attached. These requirements have been met without using shock and vibration mounts, which are undesirable in test sets because they increase size and weight.

RANGE CAPABILITY

The range capability of a radar, like that of any radio system, depends upon three things; the transmitted power, the loss in the medium, and the minimum perceptible received signal. Two of these can be combined by taking the ratio of the radiated signal to the minimum perceptible received signal. This ratio, ordinarily expressed as a level difference in db, is variously termed the "system performance," "over-all performance" or merely the "level difference." It may be determined by separate measurement of transmitter power and receiver sensitivity, or by a single overall measurement. With the powers and sensitivities commonly employed in radar, the level difference is of the order of 150 to 180 db.

The actual range that can be spanned for a given performance ratio varies considerably. For a given transmitted power, the echo power received by a radar theoretically varies inversely as the fourth power of the range. The reason for this is simple. In free space the power intercepted by a target which is small in comparison with the area of the radar beam in its vicinity will vary according to the inverse square of the distance from the radar. Similarly, that fraction of the energy reflected from the target which is intercepted by the receiving antenna will vary as the inverse square law. Since the received power involves the product of these two factors, the relation becomes:

$$P_r = K \frac{P_t}{R^4}, \quad \text{or}, \quad R = \left(K \frac{P_t}{P_r} \right)^{\frac{1}{4}} \tag{1}$$

where P_t and P_r represent, respectively, the transmitted and received power, R the range and K a constant determined by antenna design, character of target, etc.

Under operating conditions considerable departure from the above relationship may be experienced, due to such factors as (1) the curvature of the earth, (2) interference between the direct beam and single or multiple reflections, and (3) attenuation due to atmospheric absorption. Except under

conditions of severe attenuation such as may occur at the very short wave-
lengths, the received power commonly varies somewhere between the inverse
fourth power and the inverse 16th power of distance.   To state it another
way, the change in effective range is somewhere between the fourth and
the sixteenth root of the change in system performance.   The former condi-
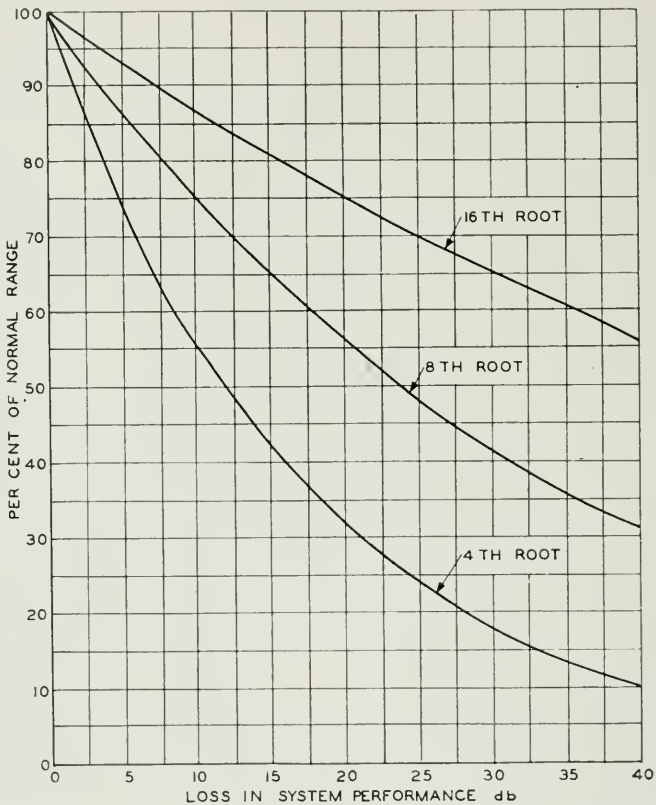
Fig. 4—Effect of reduction in system performance on radar range.

tion might hold for high angle plane-to-ship search in clear weather, the
latter for ship-to-ship search in fog.   The loss of range resulting from a given
degradation of system performance is shown for 4th, 8th and 16th root laws
by the curves of Fig. 4.   Whatever the propagation law, reduction in per-
formance always means loss of range.

Field surveys have shown that when test equipment is not available or
not used, radars in the field are likely to give no more than $\frac{1}{2}$ to $\frac{1}{4}$ of the
maximum range of which they are capable.   Hence it is necessary to know

with good accuracy the over-all performance. Known, or so-called standard, targets have often been used in the field for checking performance. Because of wide variations in transmission due to many factors, results obtained from such targets are frequently misleading.

SIGNAL GENERATORS

Signal generators for radar work deliver one or more different types of test signals, which may serve a variety of functions. More important among these functions are tuning or alignment of the radar components (TR and RT boxes, converter, beat oscillator, AFC, etc.), measurement of receiver sensitivity, checking TR and receiver recovery, measurement of loss, detection of frequency pulling, check of AFC following, measurement of IF bandwidth, check of automatic range tracking, measurement of standing wave ratio and check of video "gating" circuits. Many signal generators include means for measuring the power and frequency of the test signal, and also of an incoming signal.

Types of Signals

The test signals delivered by a signal generator may be CW, pulsed, or frequency-modulated. Occasionally square wave or sine wave modulation is provided.

Pulsed signal generators deliver a succession of single RF pulses or pulse trains, either of these generally synchronous with the pulses of the radar under test. Multivibrator or trigger techniques[1] are used to generate the pulses for modulating the microwave generator. The trigger pulse for synchronizing the pulsing circuits is commonly produced by rectifying RF pulses from the radar transmitter, thus avoiding a separate video connection to the radar. To avoid possible difficulties in video response, the RF test pulses should be of comparable width to those of the radar under test. For observing the test signals, either the radar indicator or an auxiliary oscilloscope may be used. With the single pulse method, provision is usually made for varying the delay of the test pulse with respect to the radar pulse. The width of the test pulse may also be adjustable or variable.

If the frequency of the signal generator is swept over a sufficiently wide frequency band, the IF output of the radar traces the curve of IF selectivity, thus producing a kind of pulse. With a suitable rate of frequency sweep, this pulse becomes comparable in width to the transmitter pulse, and when synchronized with the radar it can be used for test purposes. Since the pulse is produced in the radar, comparison of the shapes of receiver input and output pulses is not possible. The nominal duration of the pulse in the IF output is

$$T = B/\gamma \qquad (2)$$

where B is the width of IF band in cycles per second (for this purpose conveniently measured between 6 db points) and γ is the speed of frequency sweep in cycles per second per second. For best results this nominal width should be similar to the width of transmitter pulse for which the IF and video circuits are designed. This means that the scanning speed should be in the neighborhood of $B^2/2$.

A CW input of the same frequency as the transmitter produces in the output of the IF detector merely a direct current to which the video amplifier and radar indicator do not respond. However, CW test signals may be utilized by observing on a d.c. meter (built into the radar or separate) the change produced in detector current or converter current.

Receiver Sensitivity

Just as in radio, the sensitivity of a radar receiver is defined as the minimum received signal that is perceptible in the presence of set noise.* At microwave frequencies atmospheric disturbances are usually negligible, so that unless accidental or deliberate interfering signals are present, the operating sensitivity is the same as the intrinsic sensitivity of the receiver.

Receiver sensitivity is commonly stated as the minimum perceptible signal power in db referred to a milliwatt, (abbreviated as dbm). In practice, the receiver sensitivity depends upon the noise figure of the converter, the conversion loss, and the noise figure of the IF amplifier. If an RF amplifier is used, as is the practice at lower microwave frequencies, its noise figure is likely to be controlling. By noise figure in each case is meant the noise power in comparison with the thermal noise. The thermal noise in watts delivered to a load is kTB, where k is Boltzmann's constant, T is absolute temperature in degrees K, and B is the frequency bandwidth. Thus for a 4 mc band at 25° C the thermal noise is -108 dbm. With good design the over-all receiver noise is of the order of 10 to 15 db higher than thermal noise. The minimum detectable signal is usually not equal to the receiver noise but depends on the type of indicator, particularly on whether the presence of an echo is indicated by spot deflection or spot modulation.

With a CW signal generator, receiver sensitivity is measured by determining the minimum input power necessary to produce a perceptible change in meter reading. This affords a satisfactory relative measure of receiver performance, but since the radar indicator usually permits better visual discrimination against noise, the minimum input as read with the meter ordinarily differs from the minimum pulse input for barely discernible indicator response.

* The term noise is commonly used even though the disturbances are observed on a cathode-ray screen.

RF Oscillators

Beat oscillator tubes for radars deliver (with sufficient decoupling or isolation to prevent undue frequency pulling) a power of the order of milliwatts. This power being adequate for most test purposes, such tubes are well adapted for use as signal generators.

Throughout the greater part of the microwave range, reflex velocity-modulated tubes,[2] both the type with built-in cavity (Pierce-Shepherd) tuned mechanically or thermally, and that with external cavity (McNally) tuned by plugs, vanes or adjustment of dimensions, have been used. The former is more convenient for general use but the latter usually permits wider frequency coverage. Oscillation occurs when the repeller voltage is adjusted so that the round trip transit time corresponds to an odd number of quarter wavelengths. Ordinarily there are several different ranges of repeller voltage, corresponding to different numbers of quarter waves, each of which supports oscillation over a range of frequencies, called a mode of oscillation. Pulsing or frequency modulation is accomplished by applying a pulse or sawtooth wave to the repeller.

At the longer microwaves, a triode with closely spaced electrodes, or so-called "lighthouse" tube[3], has been employed in a tuned-plate tuned-grid oscillator of the positive grid type. Two coaxial lines, conveniently placed one inside the other, provide the tuning, with the feedback through interelectrode capacitances supplemented by loop coupling. The inner cavity (between plate and grid) controls the frequency of oscillation, while the outer cavity (between grid and cathode) provides a suitably high grid impedance. Mechanical arrangements are provided for tracking the tuning of the two cavities over a wide frequency range.

Some Design Principles

Standard signal generators which have been employed in the past for measuring the sensitivity of radio receivers usually deliver a known voltage across a low impedance. This voltage is applied in series with a dummy antenna to the receiver under test. In the microwave range this technique is inconvenient, and signal generators are designed to deliver test power on a matched impedance basis. Receiver sensitivity is stated in terms of power (dbm) instead of volts.

The components of a signal generator or other test unit are commonly arranged along a microwave transmission line. The wave guide type of line possesses certain advantages over a coaxial line in affording a lower loss, facilitating attenuator design as discussed in a subsequent section, etc. Hence the wave guide type of line is used in test equipment for those wavelengths where its size is not excessive, i.e. from about 4,000 mc upwards, and coaxial line for lower frequencies.
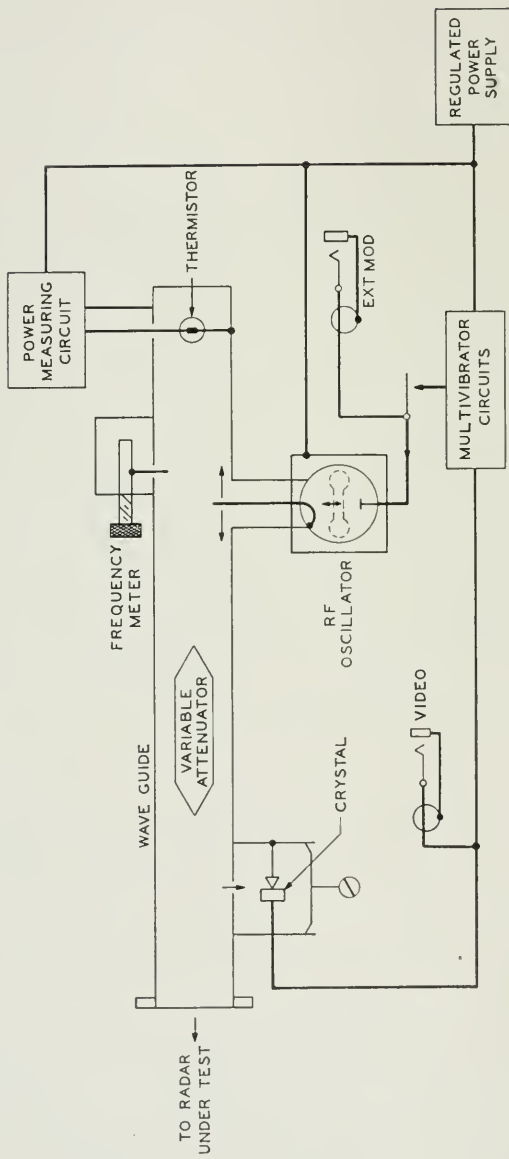
Fig. 5—Block diagram of TS-35A/AP signal generator.

A major problem in the design of microwave signal generators is the provision of shielding adequate to reduce leakage signals due to unwanted couplings or stray fields well below the minimum signal required for receiver testing. This minimum level may be as low as -70 or -80 dbm, depending on the coupling loss in the test connection.

A Pulsed and FM Signal Generator

To illustrate the functioning of a signal generator, there is shown in Fig. 5 a block schematic of a design (Army-Navy type TS-35A/AP) which covers a 12% frequency band in the vicinity of 9000 mc. An RF connection to the radar is established with a wave guide flange coupling. The frequency and power of the radar transmitter are measured by means of a coaxial-type frequency meter and thermistor power measuring circuit as described in subsequent sections. The attenuator and pad are adjusted to reduce the incoming average power to about 1 milliwatt, which gives a suitable deflection on the indicating meter. The thermistor is mounted across the wave guide.

An RF pulse train is employed in many tests. To produce this the RF oscillator output is modulated by a multi-vibrator which pulses continuously except when being synchronized. Synchronizing pulses are derived by crystal rectification of the RF pulses from the radar transmitter. The result is an initial RF pulse of 7 microseconds followed by an off period of about 10 microseconds followed by a train of RF pulses each 2 microseconds wide and recurring every 8 microseconds until resynchronization occurs at the next radar pulse.

Using the pulse train, the radar system components can be tuned for maximum sensitivity by maximizing the signal on the indicator. To check receiver sensitivity the CW power is first adjusted so that a power of 1 milliwatt is delivered to the pad and attenuator. Then with the set in the pulsed condition the amplitude of the test signal is adjusted by means of the attenuator and pad until the signal on the radar indicator is barely discernible. It is necessary for this test that the frequency of the test signal be equal to the magnetron frequency. The frequency meter is provided as part of the signal generator for this purpose.

The receiver recovery, i.e. the time required by the receiver to recover after disablement by the transmitter pulse, determines the minimum range at which a radar can be used. With this test set the receiver recovery characteristic is indicated by the amplitude of the test pulses in the interval immediately following the transmitter pulse.

The set is also adapted to serve as an FM signal generator. A sawtooth wave applied to the repeller gives a succession of frequency sweeps, each

Fig. 6—1942 model signal generator compared with one produced in 1945.

about 20 mc wide, and lasting about 6 microseconds. With this frequency modulated signal the width of receiver response may be observed on a Class A oscilloscope (i.e. one showing signal amplitude vs. time). However, with non-adjustable IF strips such measurement is seldom required. Failure of the radar AFC to follow frequency changes due to antenna scanning or other causes is indicated by a change in the indicator presentation. Pulling of the magnetron frequency due to changes in load impedance can be detected by turning off the AFC.

Signal Generator Designs

Designs of signal generators developed for the military arms during the war are interesting as landmarks of progress. The signal generator of the IE30 test set, deliveries of which began in May 1942, delivered pulsed RF signals in the 10 cm range, using sine wave synchronization. Following only three months later was the signal generator of the Army IE57A and Navy LZ test sets (Fig. 6), which covered a then very broad frequency band of 20% in the vicinity of 10 cm, and was designed to be triggered by the incoming RF pulse from the radar instead of by a separate synchronizing connection. This set and a redesigned version of it have seen wide usage in testing Army, Navy and Marine Corps radars.

Delivery of a test set for the 3 cm range, designated TS-35/AP, started in the fall of 1943. This set furnished both a train of pulses and a train of FM signals, both of which features have proved valuable. It covered a 9% frequency band with no tuning adjustment except for the oscillator. An improved design known as TS-35A (see Fig. 6) covered a 12% band.

Progress in reducing the size and weight of the test units is indicated by the fact that the IE30 signal generator weighed 121 lbs., IE57 74 lbs., whereas TS-35 and TS-35A weigh approximately 30 lbs.

FREQUENCY MEASUREMENT

Usually a radar need not operate at a precise frequency. Accurate measurements are required in the field, however, to keep the operating frequency within limits, to set the local oscillator, to check the measuring frequency, etc. In the laboratory, accurate frequency measurement is fundamental.

Frequency measurement in the microwave range is ordinarily accomplished by (1) a resonant coaxial line or (2) a resonant cavity, generally cylindrical. These types are illustrated in Fig. 7. Sometimes a combination of the two, referred to as a hybrid or transition type resonator, is employed. The measurement is actually one of wavelength, with the scale calibrated in frequency or a conversion chart provided. Some specific designs of frequency meters are shown in Fig. 8.
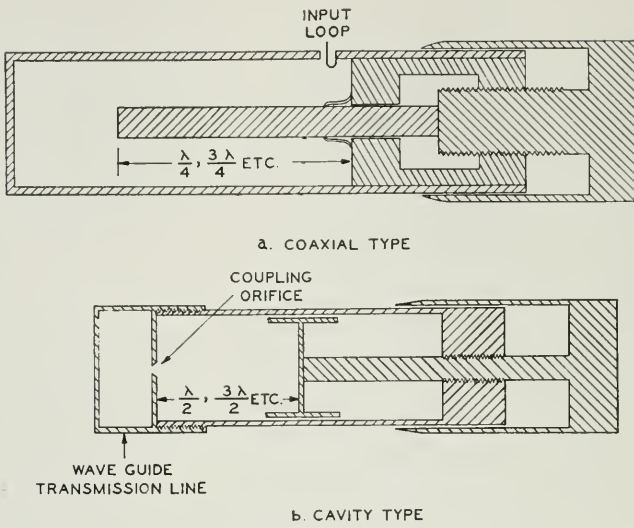
Fig. 7—Types of frequency meters.
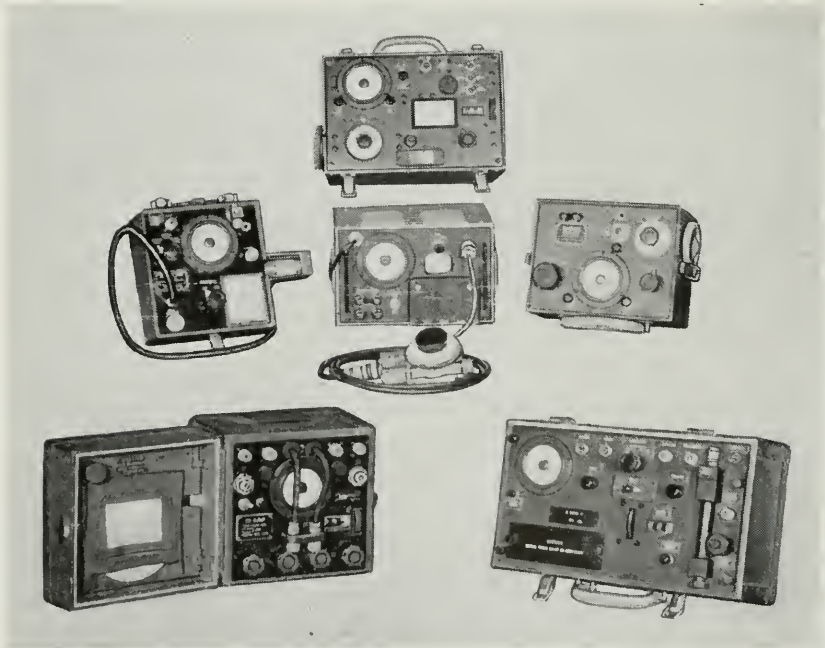
Fig. 8—A group of frequency and power meters designed for various bands in the frequency range 500 to 25,000 mc.

Coaxial Wavemeters

A coaxial wavemeter is formed of a section of coaxial transmission line of small enough diameter so that only the coaxial mode (in wave guide notation $TM_{0,0,n}$) can exist. Usually the line is short-circuited at one end and open at the other, in which case resonance occurs at the odd quarter wavelengths $\lambda/4$, $3\lambda/4$ etc.). The open circuit is obtained merely by terminating the inner coaxial conductor, the continuing outer conductor acting as a wave guide below cutoff. Sometimes the line is short-circuited at both ends, giving resonance at the even quarter-wavelengths ($\lambda/2$, λ, etc.).

Cylindrical Cavity Wavemeters

A cylindrical cavity wavemeter is merely a section of cylindrical wave guide transmission line[4] whose length is varied. In order to avoid confusion with other modes, it is preferable to use the dominant mode (described in wave guide notation, $TE_{1,1,n}$*) i.e. the mode with the lowest cutoff frequency. The cutoff wavelength (λ_c) of this mode is $1.706D$, where D is the diameter in meters. For a higher Q it may be necessary to use the circular electric mode $TE_{0,1,n}$. The cutoff wavelength for this mode is $.82D$. No useful purpose is served by using modes with l and m subscripts above unity. TM modes are often used for fixed frequency cavities, but for variable cavities TE modes are preferable since these have zero current at the inner wall of the cylinder and thus obviate moving contact difficulty. If any mode higher than the dominant one is used, suppression of unwanted modes may be required.

The accuracy of a wavemeter is dependent on its resolving power. This in turn depends upon Q, which is an index of the decrement of the resonant circuit, and is equal to $f/\Delta f$, where Δf is the distance between 3 db points on the resonance curve.

In a coaxial wavemeter, maximum Q for a given inner diameter is obtained with a diameter ratio of about 3.6[5] The basic Q of a coaxial wavemeter, assuming copper of standard conductivity, is roughly[6]

$$Q_0 = 0.042D \sqrt{f} \tag{3}$$

This expression neglects end effects and hence gives somewhat too high a value of Q.

The basic Q's for $TE_{1,1,n}$ and $TE_{0,1,n}$ cylindrical cavity resonators employing copper of standard conductivity are, respectively,

---

* TE and TM represent, respectively, transverse electric and transverse magnetic. The subscripts l, m, n denote, respectively, the number of wavelengths around any concentric circle in the cross section, the number of wavelengths across a diameter, and the number of half wavelengths along the length of the cylinder.

$$Q_0 = 0.0937 \times 10^{10} \frac{A^3}{\sqrt{f}} \frac{1}{1 + B^2\left(0.826 \frac{B}{n} + 0.295\right)} \tag{4}$$

$$Q_0 = 0.2762 \times 10^{10} \frac{A^3}{\sqrt{f}} \frac{1}{1 + B^2\left(2.439 \frac{B}{n}\right)} \tag{5}$$

with $A = \frac{\lambda_c}{\lambda}$, $B^2 = A^2 - 1$, and f is in cycles per second.

The value of Q which determines accuracy is not the basic Q, but the loaded or working Q, herein designated Q_L.

The resolving power of a wavemeter used for measuring a single frequency can be made considerably better than f/Q_L. With a sensitive meter it is readily possible to detect differences less than 1 db, which corresponds to a frequency interval of $f/2Q_L$.

The required accuracy in a wavemeter is generally absolute rather than a percentage. Hence increasingly large values of Q_L are required at the higher frequencies. Thus for a resolution of 1 mc, assuming 1 db discrimination, the values of Q_L required for different frequencies are:

Frequency	Q_L	Frequency	Q_L
1,000 mc	500	10,000 mc	5,000
3,000 mc	1,500	25,000 mc	12,500

An unnecessarily high value of Q_L has the disadvantage of making it more difficult to find the desired frequency.

Linearity

The displacement of the coaxial plunger of a coaxial type wavemeter for resonance is substantially a direct linear function of free space wavelength and if an ordinary centimeter micrometer drive is used it is possible to read wavelength differentials directly. Over bandwidths less than 20 per cent, displacement vs. frequency is also quite linear which is of considerable advantage for some uses.

For the cavity type wavemeter the displacement is a variable function of free space wavelength and becomes very non-linear as the cutoff frequency of the guide or cavity is approached. This is evident from the relation between wavelength in the guide, λ_g, wavelength in free space, λ, and cutoff wavelength, λ_c:

$$\lambda_g = \frac{\lambda}{\sqrt{1 - \frac{\lambda^2}{\lambda_c^2}}} \tag{6}$$

A cam or mechanical linkage may be employed to obtain a linear scale.

Frequency Coverage

Increasing the length of a wavemeter is desirable because this gives a larger mechanical displacement for a given frequency interval. The permissible increase is limited, however, by ambiguity with the next lower mode in going toward the upper end of the frequency scale and with the next higher mode in going toward the lower end. This means that, if ambiguity is to be avoided, the ratio of top to bottom frequency cannot exceed $(n + 2)/n$ for a coaxial line of n quarterwaves. For a cylindrical cavity resonator the ratio for the $0,1,n$ mode must be less than $n + 1)/n$, the exact limit depending on proximity to cutoff.

Guideposts

The following guideposts are suggested for choosing the type of wavemeter in the microwave range. Where limitations of size and Q permit, the coaxial quarter-wave type should be used because of its greater linearity. If this type is inapplicable, the cavity type with $TE_{1,1,n}$ should be used unless its Q is inadequate, in which case $TE_{0,1,n}$ should be employed.

Couplings

A loop, orifice or probe may be used for coupling to a wavemeter. Coupling to a coaxial wavemeter is generally effected by a loop placed near a short-circuited end so as to be in the maximum magnetic field. For coupling to a cylindrical cavity wavemeter, an orifice in or near the base of the cavity is usually employed. The coupling to the wavemeter is kept small enough to avoid serious reduction of loaded Q.

Types of Detectors

Various types of detectors may be associated with a wavemeter, the most commonly used being (1) a crystal rectifier and microammeter, or (2) a thermistor bolometer. When a crystal rectifier is employed with a cavity or coaxial wavemeter a circuit similar to that shown in Fig. 9 is used. Important items in such a circuit are the "RF by-pass" condenser, and the "video" condenser. The latter, by providing a low-impedance path to the video signals, improves rectification efficiency when the input signal is pulsed. The quarter-wave stub shown in the figure is used when the input or coupling circuit does not provide DC and video paths. When the signal is pulsed at a low duty cycle, high peak currents through the crystal are obtained even though the average current through the meter is small, and it is possible to impair or burn out the crystal unless extreme care is taken. The use of a thermistor, which is self-protecting for large overloads, avoids this danger. Another expedient is to limit the crystal current to a small value and employ a video amplifier and oscilloscope.

Methods of Use

A wavemeter may be used as either a transmission or a reaction instrument. In the former case (Fig. 10a) it is inserted directly in the transmission path, so that substantially no through transmission occurs except at resonance. In the latter case (Fig. 10b) the wavemeter is coupled to the transmission path or a branch circuit. When the meter is tuned off resonance it presents such a high impedance to the main path that its effect is negligible. At resonance, however, it offers a lower impedance which reflects energy in the main line so that less power reaches the detector and
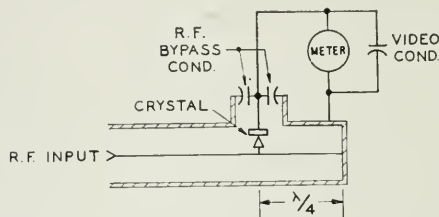
Fig. 9—Crystal detector for pulsed RF signals.
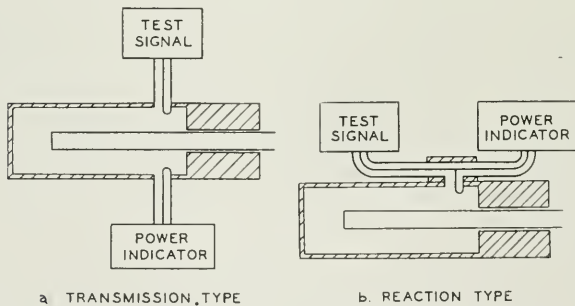
Fig. 10—Types of frequency meter circuits.

a dip in the reading occurs. For most applications the reaction arrangement is preferable since the power transmitted when the wavemeter is off tune may serve various purposes. For analysis of frequency spectrum the transmission method is necessary. This method requires two couplings to the wavemeter, which lowers the Q as compared with the reaction type.

Drive and Scale

A direct drive with a precision lead screw of the micrometer type is frequently used. Accuracy of reading is insured by spring loading to minimize backlash and by specifying close tolerances for threads and for concentricity of plunger and cavity. The scale on a wavemeter must be fine enough to permit utilization of the resolving power. The conventional micrometer

type of scale sometimes makes accurate reading difficult in a small compact meter. Scale mechanisms which have been used include counter types, clock face or expanded drum types with gearing between vernier and coarse scales, and a single direct reading scale with divisions arranged in a spiral for compactness.

Effect of Temperature and Humidity

To minimize the effects of temperature on the accuracy of readings, invar has been employed for elements whose dimensions affect the wavelength. For accurate work, the scale reading must be corrected for temperature. A rough approximation is that the scale reading varies in accordance with the coefficient of expansion of the metal.

Water vapor included in the air dielectric of a wavemeter has an appreciable effect on the dielectric constant and hence on the resonant frequency.[7] Thus, for example, in going at sea level from 25° C, and 60% humidity to 50° C, and 90% humidity, the scale reading should be reduced by .03%. Correction can be made by means of a chart, a convenient form of which has been prepared by Radiation Laboratory.

Calibration

Frequency meters are calibrated against sub-standards which in turn are calibrated against a multiplier from lower frequencies for which a high order of accuracy can be obtained. Such multipliers have been made available at Radiation Laboratory and the National Bureau of Standards. The accuracy obtained at interpolation frequencies is of course less than at exact multiples of the base frequency. In the microwave range the accuracy is believed to be of the order of one part in 100,000.

POWER MEASUREMENT

There are two needs for power measurement in radar maintenance, namely (1) in evaluating transmitter performance and (2) in standardizing test signals. Power output is, of course, only one factor in transmitter performance, others being (a) frequency and (b) spectral distribution or shape of RF envelope. Ability to measure absolute power is desirable to permit interchangeable use of test sets in the field.

Measurement of Pulse Power

The transmitter power as used in Formula (1) is the average power during the pulse. The relationship of pulse to (long) average power is

$$\frac{P_t \text{ av.}}{P_t \text{ pulse}} = Tf_r \tag{7}$$

T is the pulse duration in seconds and f_r is the pulse recurrence frequency (P. R. F.) in cycles per second. The product Tf_r is the duty cycle. (Sometimes the reciprocal of this number is referred to as the duty cycle. The magnitude is usually such that no ambiguity arises.)

During the early days of radar it was the practice to measure pulse power. The test equipment was coupled to the radar by a path of known loss. The RF envelope was derived by means of a crystal rectifier and applied to an oscilloscope. With the aid of an RF attenuator the level applied to the crystal rectifier and oscilloscope was held constant. Calibration was obtained by using a signal generator whose output was standardized, prior to pulsing, with an averaging type of power meter. The procedure was rather involved, with several sources of possible error. Since it is much simpler to measure (long) average power, field measurement of pulse power was soon abandoned. Though the pulse power can be computed from average power if the pulse width, pulse shape and repetition rate are known, it soon became the practice to specify field performance requirements in terms of average power.

Thermistor Power Meters

A number of devices have been used for measuring average power in the microwave range. Those suitable for handling the small amounts of power normally involved in field tests include (1) thermistors, (2) platinum wires and (3) thermocouples. In each case the RF power to be measured is absorbed in the measuring element. The measurement consists in observing the resistance change in the thermistor or platinum wire, or the thermoelectric voltage from the thermocouple. By analogy with devices used for measuring minute quantities of radiant heat, either a thermistor or a platinum wire instrument is sometimes referred to as a bolometer. The platinum wire device has also been termed a barretter.

A thermistor for microwave power measurement is a tiny bead (about 5 mils in diameter) composed of a mixture of oxides of manganese, cobalt, nickel and copper, constituting a resistor with a very high negative temperature coefficient.[8]

The thermistor has a number of advantages for microwave work, namely: (1) resistance is highly sensitive to change of heating power, which obviates any need for amplification or a super-sensitive meter, and makes it possible to use a rugged d.c. meter; (2) reactance is low compared with RF resistance, which makes it possible to incorporate the thermistor in a power absorbing termination which matches the impedance of a microwave transmission line over a wide band; (3) resistance change is the same function of electrical heating power at any frequency, which permits direct comparison of the unknown microwave power with easily measurable d.c. power; (4) sensitivity

to damage and burn-out is inherently low, and added protection results from impedance mismatch during overload. Because of these characteristics thermistors have been far more widely used than other detectors for microwave power measurement. Broad-band thermistor mounts have been designed to match both wave guide and coaxial transmission lines, the latter not only in the microwave range but also down to low frequencies. Some of the test sets specifically intended for power measurement or for combined power and frequency measurement are shown in Fig. 8.

The change in the thermistor resistance due to RF heating current is determined by placing the thermistor in one arm of a d.c. bridge. By noting the d.c. power necessary to balance the bridge with and without RF power in the thermistor, the magnitude of the RF power may be determined. For most purposes, however, a direct reading power meter is preferable. This can be obtained over a moderate range of power levels by employing an unbalanced bridge. The bridge is balanced for d.c. only and the measurement consists in noting the meter deflection when RF power is added.

The resistance of a thermistor is a highly sensitive function not only of electrical heating power but also of ambient temperature. For convenient field measurement, the effect of ambient temperature must be cancelled out in the indicator circuit so that the indication depends only on RF power.

Water Loads

A method which has been used in the laboratory and factory for measuring high-level microwave power consists in terminating the RF transmission line in a water load arranged as a continuous flow calorimeter. This method can be made quite accurate but is cumbersome. More recent practice is to terminate the RF line in a solid load of a type described later in this article, and to couple a thermistor power meter to the line by means of a directional coupler (described below) of known loss. Very close correlations have been obtained between the two methods over the entire microwave band.

Echo Boxes

A device unique to radar testing is a high Q resonant cavity, known as an "echo box" or "ring box." The cavity is coupled to the radar transmission line or antenna as indicated in Fig. 11. During the transmitted pulse, microwave energy is stored in the cavity. In the period immediately thereafter, energy is returned to the radar over the same path, producing a signal on the radar indicator. The energy in the cavity builds up exponentially to an amplitude dependent on the radar power. At the end of the pulse the returned energy decays exponentially, disappearing into the noise at a point determined by receiver sensitivity. The time interval between the

end of the transmitted pulse and the point where the signal on the radar indicator disappears into the background noise, called the "ring time," therefore measures the over-all performance of the radar.

An echo box is a particularly useful instrument for radar testing because it measures over-all performance directly, because it permits a rapid tune-up, and because it utilizes the radar transmitter as its only source of power and therefore can be made extremely portable. Figure 12 shows typical ring-time patterns on different types of indicators. In actual practice the ring-time is read in miles on the radar range scale and hence is measured from
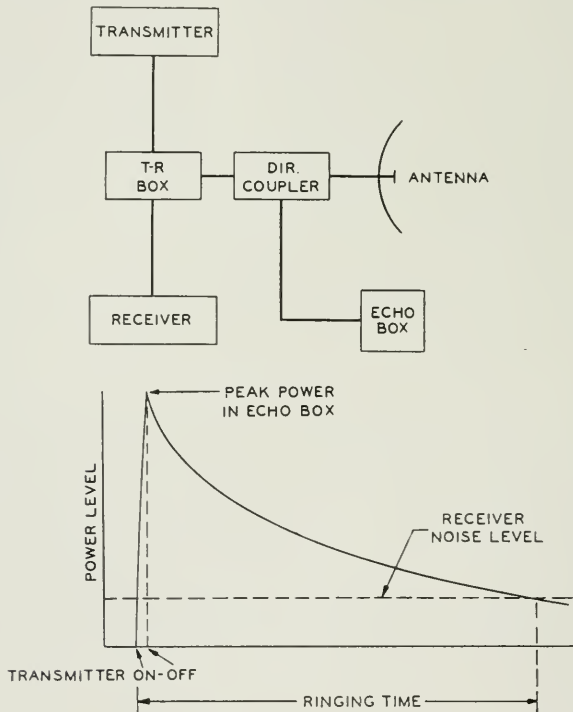
Fig. 11—Radar test with echo box.

the beginning of the transmitted pulse rather than the end. The difference is unimportant, however, since standard and limiting values of ringtime are established for operating conditions. It will be noted that the echo box does not return a true echo to the radar, so that the name "echo box" is not entirely appropriate.

Types and Uses

Echo boxes are of two general types, tuned and untuned. A tuned echo box is designed to resonate in a single mode adjustable over the operating

frequency range. An untuned echo box is a fixed cavity of a size sufficient to support a very large number of modes within the working range. Tuned echo boxes are more versatile and more widely used than untuned boxes.

The most common type of tuned echo box is designed for hand tuning. While other shapes are possible, the most convenient one is a right cylinder whose length is adjusted by a movable piston. The $TE_{0,1,n}$ mode gives
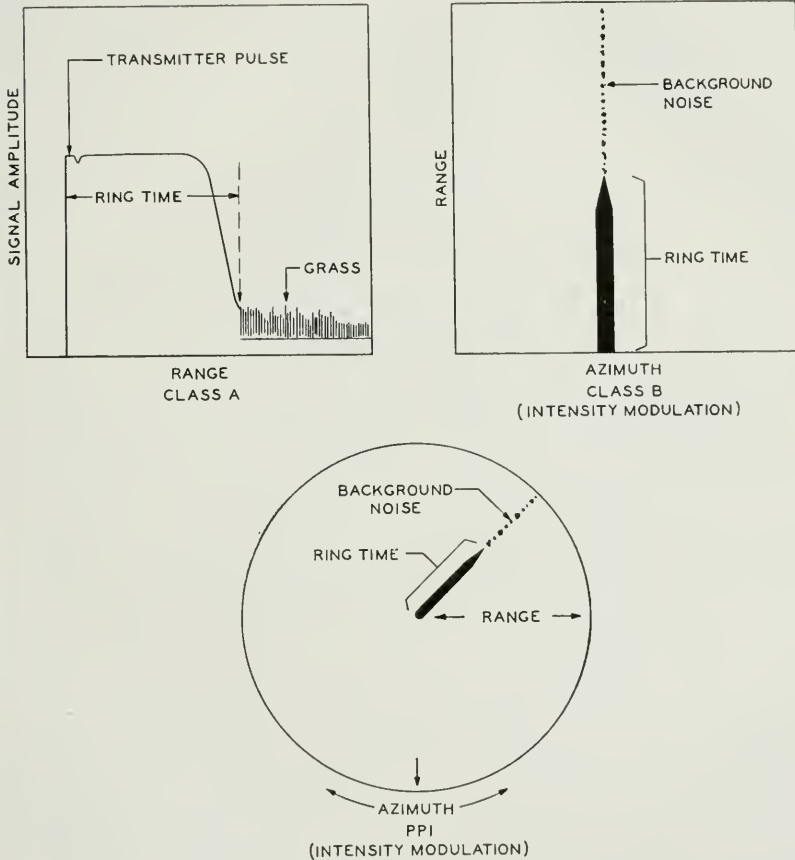
Fig. 12—Typical ringtime patterns on radar indicators.

maximum Q for a given volume and minimizes the number of unwanted modes present within the desired band. The value of n is determined by the desired value of Q (see formula 5). Unwanted modes can be partially avoided by choice of design parameters. However, for high values of Q, and especially for broad frequency bands, the suppression of unwanted modes involves design problems of the highest order.

As indicated in Fig. 13, a tuned echo box cavity is usually provided with two couplings. One of these is to the radar pick-up; the other to an attenuating device, crystal rectifier, and meter, which serve for tuning the cavity and for other purposes. With such an instrument, not only can the radar be tuned up and its over-all performance determined, but many other tests can be made, to wit: (1) the setting of the plunger at resonance indicates the transmitter frequency or wavelength; (2) calibration of the crystal affords a rough measure of output power; (3) since the Q required for adequate ringtime is so high that the cavity selects only a narrow segment of the transmitter spectrum, a spectrum analysis can be made by plotting frequency versus crystal current reading; (4) slow recovery of TR box and receiver after the transmitted pulse can be detected by noting the behavior of the ringtime pattern at short ranges as the echo box is detuned; (5)
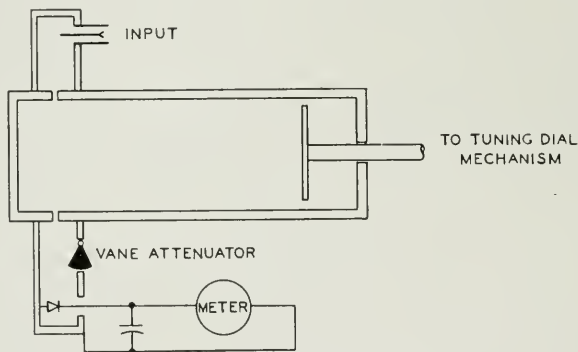
Fig. 13—Functional schematic of tuned echo box

inability of the receiver to recover promptly after a strong signal (the result of imperfect d.c. reinsertion in the video amplifier or of overloading of the I.F. amplifier) is indicated by a blank following the end of the ring; (6) improper pulsing (e.g. double moding or misfiring) can be determined with a class A oscilloscope; (7) the frequency and power of the local oscillator can be measured. In tuned echo boxes, requirements for extreme fineness of tuning control and precise resettability have given rise to interesting problems in the design of the mechanical drive and indicating mechanism.

In another type of echo box, hand tuning is supplemented by motor-driven tuning or so-called "wobbling" over a frequency range wide enough to embrace expected variations in transmitter frequency. Operation is controlled by a single push-button which energizes the motor and actuates the cavity coupling. Such an instrument may be permanently installed in a plane and used to check the radar during flight.

For an untuned or multi-resonant **echo box,** rectangular shape is convenient. The box should be large enough to make it highly probable that over the operating band one or more modes will be present within any frequency interval of width equal to the main concentration of the transmitter spectrum. For a given rectangular volume a cube gives the largest number of modes. The total number of modes up to a frequency of wave length λ is

$$N_M = 8.38 \; V/\lambda_0^3 \qquad (8)$$

where V is the volume. However, because of the cubical shape many different modes tend to coincide in wavelength, a condition referred to as
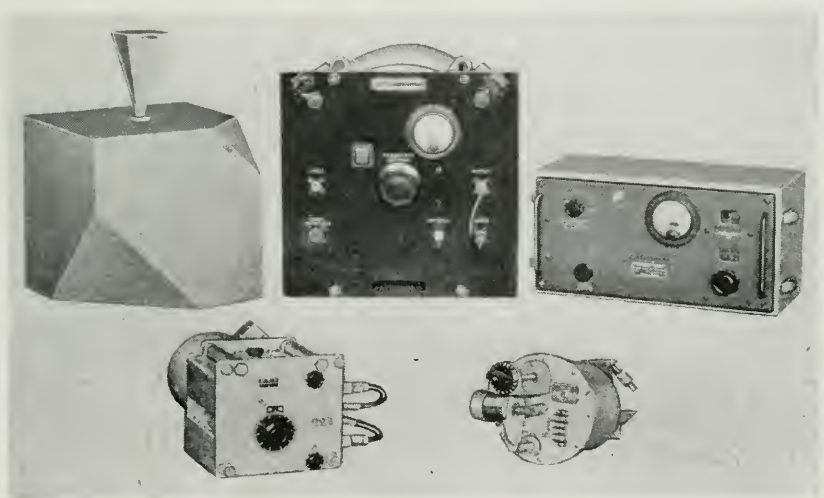
Fig. 14—A group of echo boxes of various types.

degeneracy. To spread out the modes, the box is made slightly off cube and one or more corners sliced off. At the longer microwaves the size of box is determined by the number of modes, and the size becomes quite awkward. For the shorter microwaves the size is determined by the required value of Q. Hence the use of untuned echo boxes has been limited to the frequency range from about 9000 mc upwards with sizes of the order of 12 to 24 inches on a side. Even with an extraordinarily high probability of finding modes within the radar band, substantial differences in response are found for relatively small changes in frequency. Accordingly untuned echo boxes are more useful for rough tune-up than for precise measurement.

A number of specific designs of echo boxes for different microwave bands are shown in Fig. 14.

Q and Ring Time

For satisfactory measurement the ringtime must extend beyond nearby echoes which would obscure the test signal. For most radars a ringtime of 20 to 30 microseconds (about 2 to 3 miles) has been found satisfactory although considerably higher values have sometimes been provided. Even apart from echoes, a long ringtime is desirable since this gives a lower decay rate and a more sensitive measurement.

Computation will show that an extremely high value of Q is necessary to obtain the desired ringtime. For maximum ringtime the cavity coupling should be such as to make the working Q (Q_L) about 90 per cent of the non-loaded Q. Values of working Q which have been provided in different frequency ranges are approximately as follows:

Frequency	Q_L	Frequency	Q_L
1,000 mc	70,000*	10,000 mc	100,000
3,000	40,000	24,000	200,000

* In this case a higher Q was needed for a long range ground search system.

The difference in performance corresponding to a given change in ringtime can be determined from the decay rate which is

$$d = 27.3 \, f/Q_L \, db/\text{microsecond} \qquad (9)$$

For a given frequency the ringtime is directly proportional, and the decay rate inversely proportional, to Q. For a given ringtime, the required Q is directly proportional to frequency.

Accurate measurement of extremely high Q's is essential in echo box work. A decrement method, in which a pulsed RF oscillator and oscilloscope are used to determine the loss corresponding to a known time interval, has proved most satisfactory.

SPECTRUM ANALYSIS

The frequency components of a non-repetitive rectangular d-c. pulse may be determined by well known methods using Fourier integral analysis. The envelope of amplitudes is of the form $(\sin x)/x$ where $x = \pi f T$. This envelope is shown by the right-hand side of the curve of Fig. 15a, f_0 being assumed to represent zero frequency. The first zero occurs at the frequency $f = 1/T$.

Similarly the envelope of the spectrum of a rectangular a-c. pulse is given by the complete curve of Fig. 15a, f_0 in this case being the carrier frequency. For a non-repetitive pulse all frequencies are present in amplitude as shown by the envelope. When a stable carrier frequency is pulsed at uniform intervals and in precise phase relation, only harmonics of the repetition

frequency are present under the envelope. In radar practice conditions are, as a rule, not sufficiently stable for this to occur.

Because of its bandwidth, an echo box cannot reproduce the ideal spectrum envelope of Fig. 15a. Instead the curve for a good spectrum may resemble that of Fig. 15b, while spectrum irregularities detrimental to radar performance may be revealed by curves such as those of Fig. 15c and 15d. Broadening of the spectrum is undesirable because less energy falls within the receiver band. Energy removed from the main concentration may result from double moding or from the occurrence of a different frequency during the rise or fall of the pulse. Frequency modulation due to a sloping
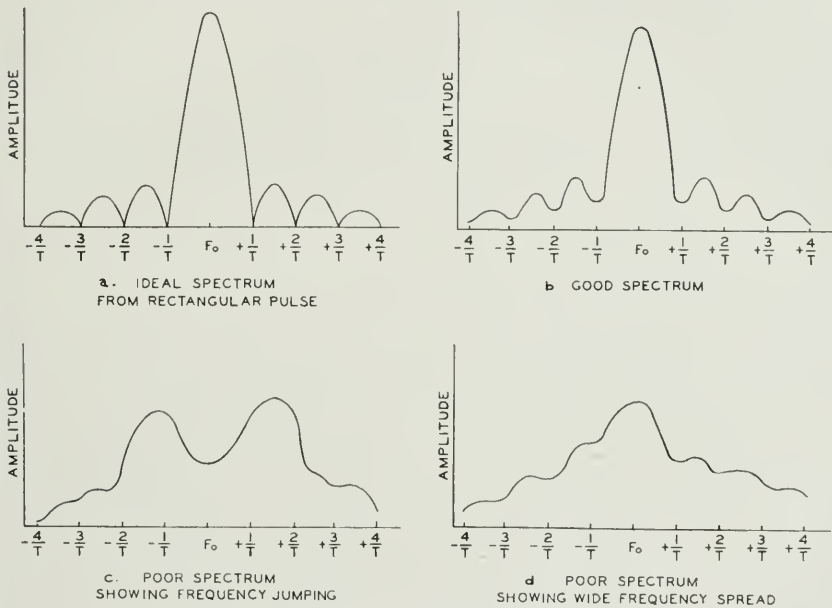
Fig. 15—Radar spectrum analysis with echo box.

or spiked input pulse produces a non-symmetrical spectrum, not infrequently characterized by a high side lobe. Frequency jump in the middle of the pulse, due to line reflection, may produce two distinct maxima.

Another device widely used for this purpose is the "Spectrum Analyzer" as developed by Radiation Laboratory, which provides an oscilloscope presentation of spectrum component amplitudes versus frequency.

STANDING WAVE MEASUREMENTS

Theory

The expression for the distribution of current or voltage along a misterminated line of appreciable electrical length yields two terms which

may be considered as representing two waves transmitted in opposite directions, one (the incident wave) from the generator toward the load, the other (the reflected wave) from the load toward the generator. The summation is a standing wave pattern. The standing wave ratio (SWR) is defined as the ratio of the wave amplitude at a maximum point (anti-node) to that at a minimum point (node). If the standing wave ratio is stated as a numeric, it is necessary to specify whether it applies to voltage (VSWR) or power (PSWR). Possibility of ambiguity is avoided by stating the ratio in db.

The ratio of the reflected current to the incident current is the *reflection coefficient*, here designated as ρ. The value of the reflection coefficient is given both in magnitude and phase by

$$\rho = \frac{Z_0 - Z}{Z_0 + Z} \qquad (10)$$

where Z_0 is the characteristic impedance of the line and Z is the load impedance. The reflection coefficient is related to the standing wave ratio as follows:

$$VSWR = \sigma = \frac{1 + \rho}{1 - \rho}, \quad \text{or,} \quad \rho = \frac{\sigma - 1}{\sigma + 1} \qquad (11)$$

Plots of the relationships are shown in Fig. 16.

The reduction of radiated power due to reflection losses in a radar transmission line, while important, is usually less serious than other effects of impedance irregularities. Since the load impedance reacts on the oscillator circuit, the frequency and output of most transmitter tubes are quite sensitive to load impedance. If the line is electrically long, so that its impedance varies rapidly with frequency, marked instability of oscillator frequency may occur, a condition referred to as "long line effect."

Since radar transmission lines contain many potential sources of impedance discontinuity, including not only the antenna but a variety of couplings, bends, wobble joints, rotating joints, switches, etc., measurements of standing wave ratio are frequently required. The need for such measurements depends in part on whether the line is "preplumbed" or is provided with field adjustments.

Devices

Standing waves may be detected and measured by several different types of devices, including (1) a slotted line, (2) a squeeze section, (3) a directional coupler and (4) a hybrid T. All of these furnish information on the magnitude of the standing wave ratio. In some cases phase information may be obtained also, which permits determination of impedance,[9] but this knowledge, while useful in the laboratory, is seldom required in field work.

A block diagram of an arrangement employing a slotted line for measuring standing waves is shown in Fig. 17. The oscillator source is commonly followed by a pad or attenuator to prevent frequency pulling. The slotted section may be either a coaxial or a wave guide line employing a mode which is not disturbed by the presence of the slot (e.g. normal coaxial mode; $TE_{1,0}$ in rectangular wave guide; $TM_{0,1}$ in round wave guide). A traveling pick-up probe or loop projects through the slot and couples energy from the line into a detector which delivers d-c. or audio-frequency to the indicator. The
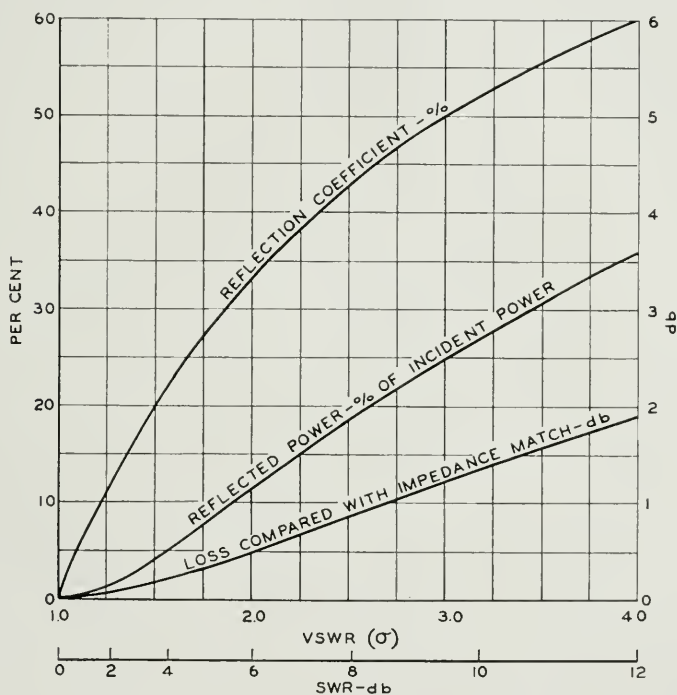
Fig. 16—Relations in mismatched transmission lines.

probe is moved longitudinally to find points of maximum and minimum field strength. To avoid distortion of the field within the line, the probe should be small and should project only a short distance inside the slot. For accurate results extreme care must be exercised in design and construction to avoid variation in depth of immersion as the probe is moved. Several slotted lines employed for standing wave measurements are shown in Fig. 18.

A squeeze section consists of a section of rectangular guide with slots milled in the center of both broad faces so that the width of the guide can be varied by external deforming means. This changes the wave length in
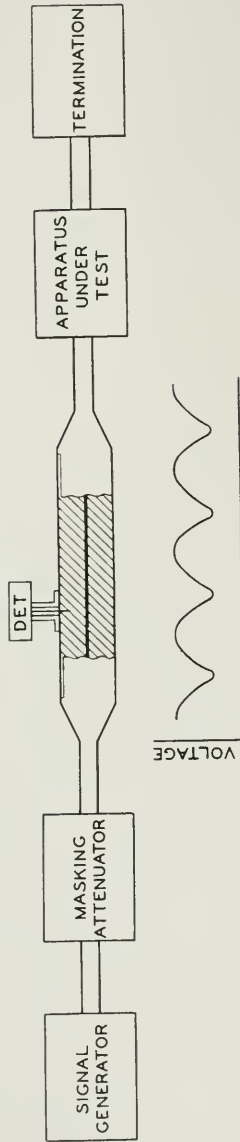
Fig. 17—Standing wave measurement.
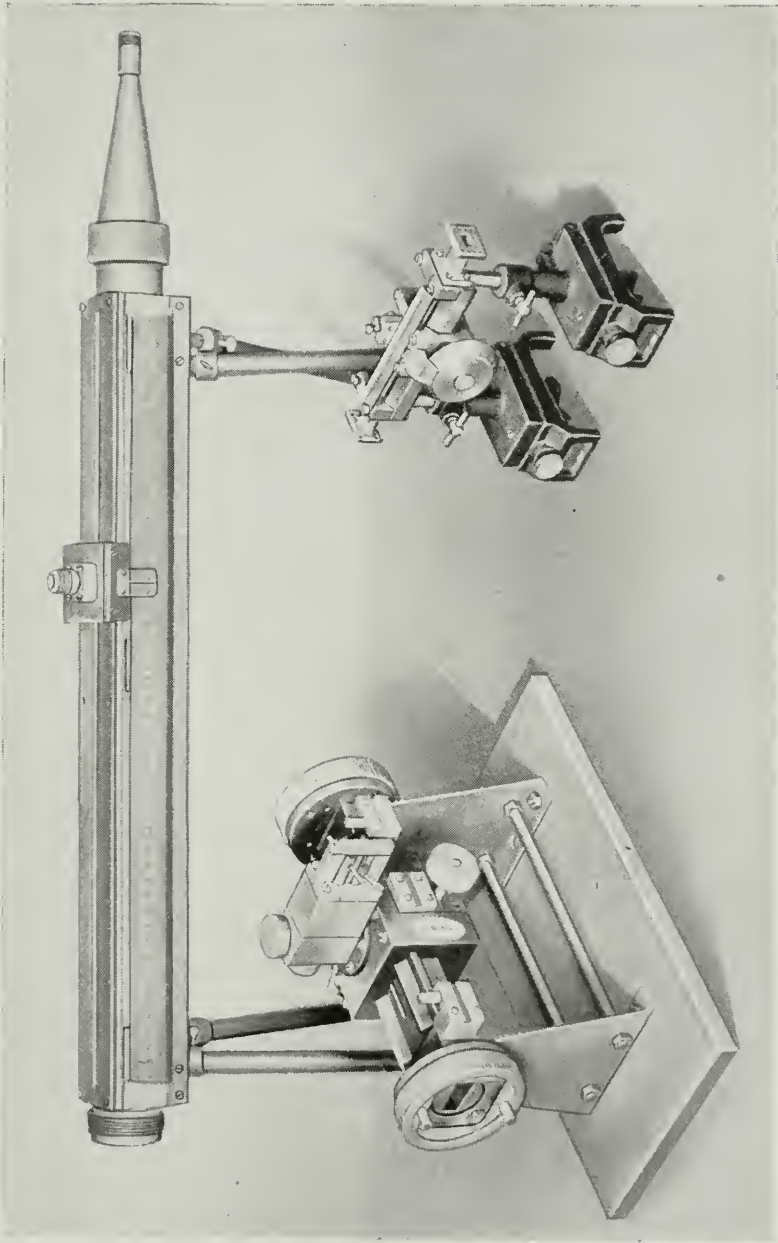
Fig. 18—Standing wave detectors used on coaxial and wave guide transmission lines.

the guide, so that maximum and minimum values may be determined by a fixed probe and indicator.

The use of a directional coupler for standing wave measurement is discussed in the following section.

Another useful device for standing wave measurement is the "hybrid T" or "magic T". This is a sort of microwave bridge, consisting of a main wave guide to which an E plane branch and an H plane branch are joined in the same physical plane. With matched terminations of the two ends of the main guide, the two branches are conjugate. Terminating one end of the main guide in the unknown impedance and the other in a matched termination, the degree of impedance mismatch of the unknown is indicated by the magnitude of the reflected wave which appears in one branch when energy is fed into the other.

Directional Couplers

Accurate measurement of transmitter power and receiver sensitivity requires a coupling path of known loss between the radar and the test set. The first method employed for this was to place a portable test antenna (see Fig. 2) in the field of the radar antenna. Depending on the frequency range, this test antenna took the form of a dipole,[10] with or without a small reflector, or an electromagnetic horn.[11] With this method it was necessary to calibrate the loss of the space coupling path between the two antennas. Since it proved difficult to locate the test antenna at exactly the same point and to be sure that the main antenna pattern remained the same, a separate calibration of the coupling loss was usually required whenever a measurement was made.

An alternative method was to place a single probe in the radar transmission line. This introduced another sort of difficulty. Accuracy of measurement was vitiated by the presence of standing waves which rendered the probe pick-up a function of frequency and of location with respect to the irregularities. A highly satisfactory answer to the entire problem was found in a device which is called a directional coupler because it couples only to the wave propagated in one direction. In its simplest form a directional coupler consists of two couplings to the main transmission line, which add for one direction of transmission and cancel for the other. Thus, for example, Fig. 10a shows a form of directional coupler for wave guide which is placed in the radar transmission line at the point indicated schematically in Fig. 1. An auxiliary wave guide is coupled to the main guide through two identical orifices spaced $\lambda_g/4$ between centers (or more generally $n\lambda_g/4$ where n is an odd integer). Assuming the incident and reflected waves in the main guide to be directed as shown, and the auxiliary guide to be terminated on

one end, a test circuit connected to the other end will be coupled to the incident wave, while theoretically the two couplings to the reflected wave will differ by $\lambda/2$ and therefore cancel one another.

With such a device measurements may be made of the characteristics of the incident wave independently of reflections. If the coupling to the main line is not too close there is no appreciable effect on the incident wave, and continuous monitoring can be had. Conversely, test signals applied through the directional coupler will travel in the main guide in the proper direction for testing the radar receiver.

If the locations of the termination and the test connection point in Fig. 19a are reversed, the couplings to the main transmission line are also reversed. Such an arrangement therefore permits measurement of the reflected power which in turn makes it possible to adjust for minimum reflected power and hence for minimum SWR. Comparison of the reflected power with the direct power determines the SWR. For convenience in measurement, two directional couplers pointed in opposite directions are frequently used, the combination being referred to as a bi-directional coupler (Fig. 19b). One advantage of this arrangement is that the ability to measure the reflected power from the antenna and that part of transmission line beyond the coupler provides means for detecting trouble in that part of the system. Directional couplers may be applied to any type of transmission line. Figure 19c shows a simple form of directional coupler for a coaxial line.

One characteristic of importance in a directional coupler is the coupling loss. A small value of coupling loss affords increased sensitivity of measurement, while a sizable value is desirable to minimize reaction on the main transmission line as well as for other reasons. A loss of around 20 db has usually been found a good compromise. It is now the practice to incorporate a directional coupler in every radar to obtain a test connection point.

Due to unavoidable imperfections, a directional coupler never gives complete cancellation for the undesired direction of transmission. The departure from ideality is indicated by the directivity (also referred to as front-to-back ratio) which is defined as the scalar ratio of the two powers measured at the test connection point when the same amount of power is applied to the main guide, first in one direction and then in the other. For measurements of the direct wave and of receiver characteristics, a moderate directivity, of the order of 15 db or better, is sufficient. In measuring reflected power, however, the directivity determines the amount of direct power which appears at the point of measurement and therefore controls accuracy. The chart of Fig. 20 will facilitate determination of the maximum
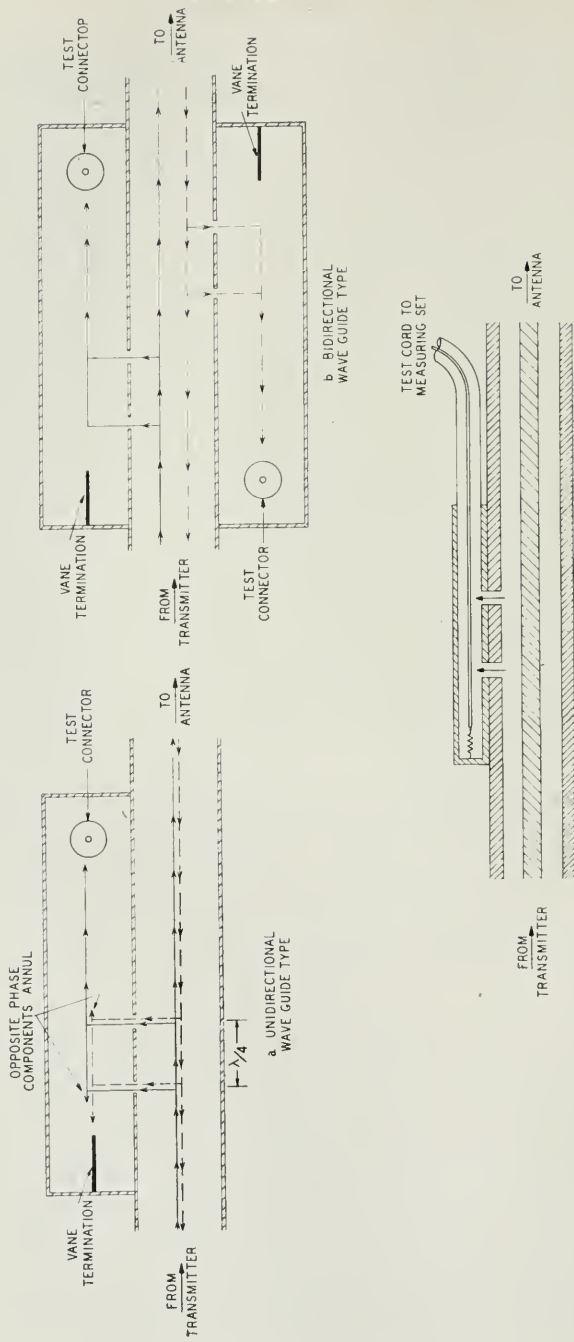
Fig. 19—Directional coupler arrangements.

error that may occur in measuring different values of SWR with various as-
sumed directivities.

With a simple two-hole coupler, the directivity deteriorates rapidly as
the frequency departs from that corresponding to quarter-wave spacing.
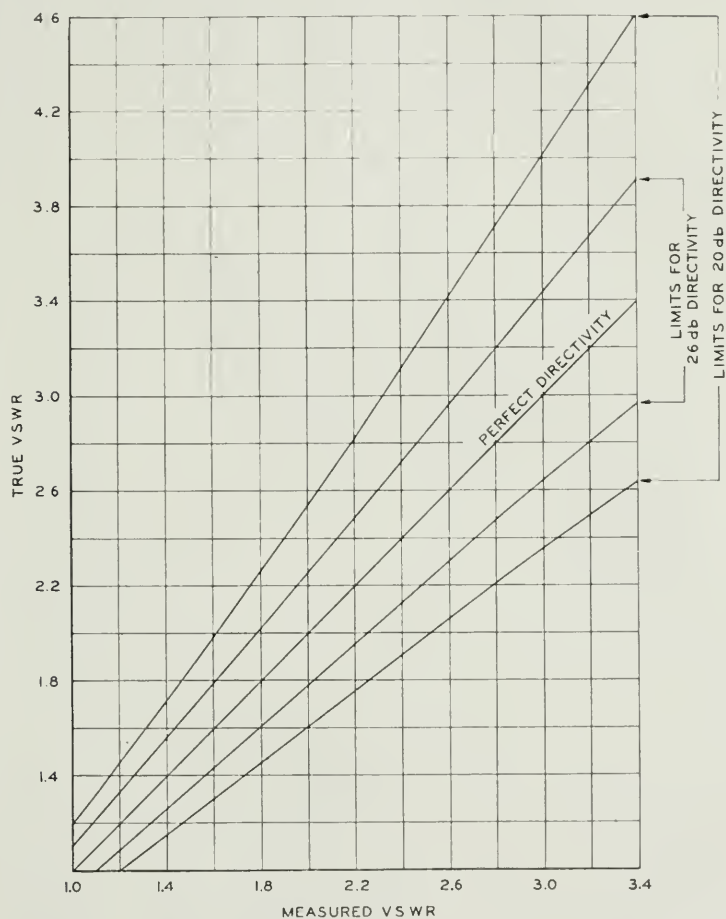
Fig. 20—Error in standing wave measurements caused by directivity of directional
coupler.

By providing additional couplings suitably spaced, the residuals from
different sets of couplings can also be cancelled against one another and the
directivity versus frequency characteristic can be materially broadened.
With multiple hole couplings a minimum directivity of 26 to 30 db over a
frequency band of 10 to 20% is readily practicable in quantity production,
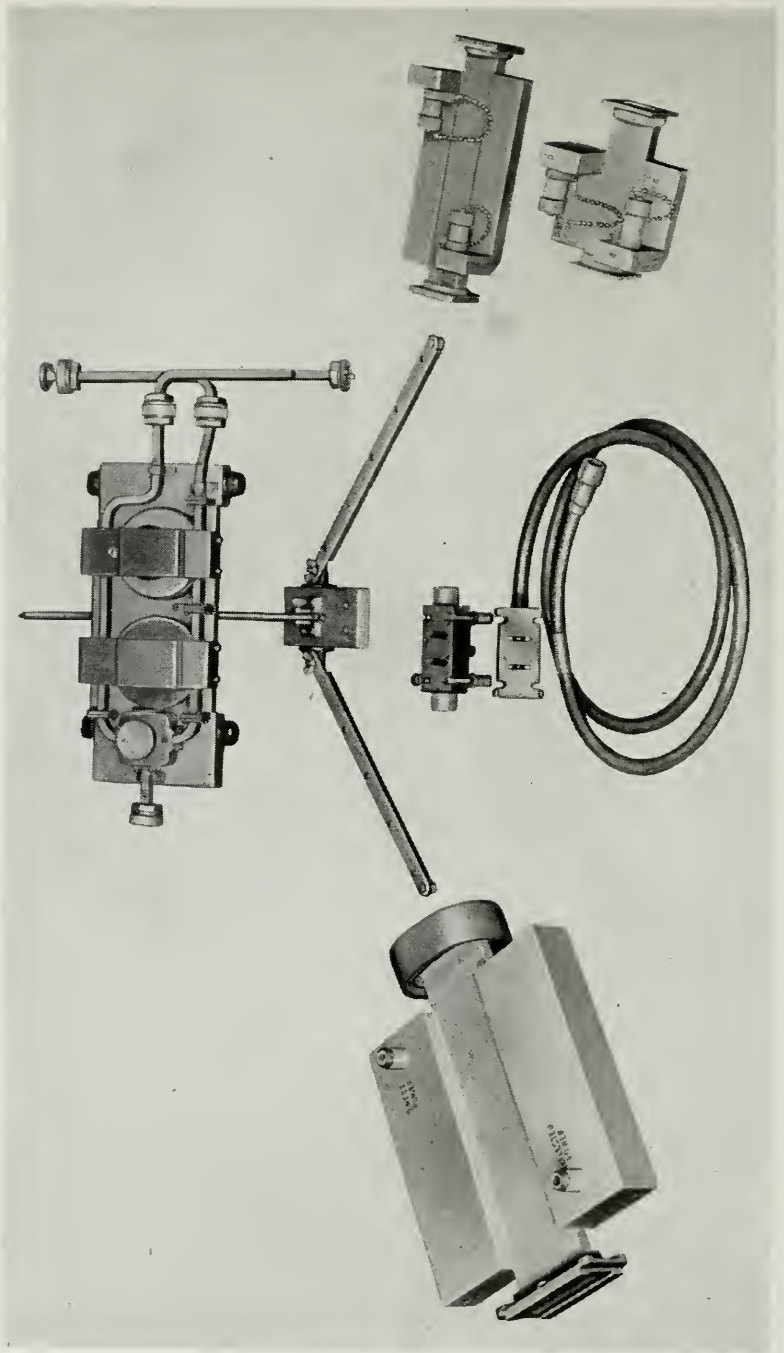
Fig. 21—A group of microwave directional couplers together with a directional coupler test set.

472

and much better values are obtained in the laboratory. Some of the numerous designs of directional couplers developed for association with operating radars are shown in Fig. 21.

Because they are more convenient than slotted lines and can be made more accurate, directional couplers have been extensively used for SWR measurement in the laboratory. A number of special arrangements have been devised to improve both accuracy and convenience. A directional coupler arrangement which has been provided for field measurement of SWR in the vicinity of 25,000 mc is also illustrated in Fig. 21. In this the direct power is brought to equality with the reflected power by an attenuator whose dial is calibrated directly in SWR. A wave guide switch facilitates the power comparison.

AUXILIARIES AND COMPONENTS

RF Loads

An RF load (or dummy antenna) which will absorb the radar power in an impedance which matches the transmission line is very useful in radar work. Such a device permits testing the radar in operating condition without actual radiation which might give information to the enemy or interfere with other radars. It also makes it possible to test the radar in locations where reflections from the ground or nearby objects would otherwise hamper or prevent a test. RF loads for microwave work usually consist of a section of transmission line (either coaxial or wave guide, depending on wavelength) containing a high-loss dielectric. The impedance of such a load is necessarily low and must be matched to the radar line by tapering the dielectric over a distance of several wave lengths.[10] Moreover, if the line is to handle high power, tapering over a considerable length is necessary to distribute the heat.

A coaxial load is preferably tapered from outer conductor to inner conductor, since this both reduces the voltage gradient and facilitates heat dissipation. A dielectric consisting of a mixture of bakelite, silica and graphite, molded in place, has been found satisfactory. For wave guides a ceramic containing carbon may be preformed, with taper in one or two dimensions, and cemented in place.

Figure 22 shows a number of RF loads developed for different frequency bands. One of these, TS-235/UP, provides an excellent impedance match over the frequency range from 500 mc to above 3,000 mc. When equipped with a blower designed for uniform transverse ventilation, it will handle a peak power of the order of 750 kw with a duty cycle of about .001.

Microwave Attenuators and Pads

RF attenuators and pads are cornerstones of microwave testing. Attenuators are used to adjust unknown signals to levels suitable for measure-
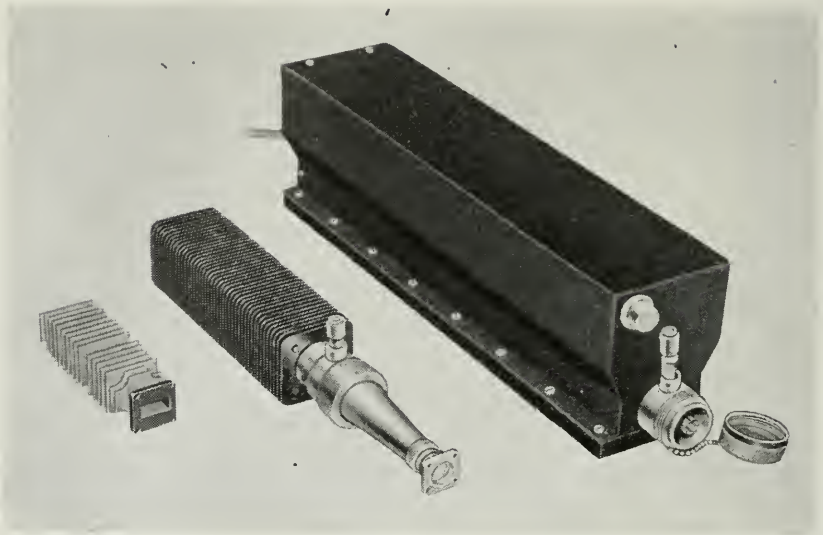
Fig. 22—RF loads for different bands in the microwave frequency range.

a. STRIP MOVING VERTICALLY IN GUIDE

TERMINATIONS

b. WAVE GUIDE BELOW CUTOFF
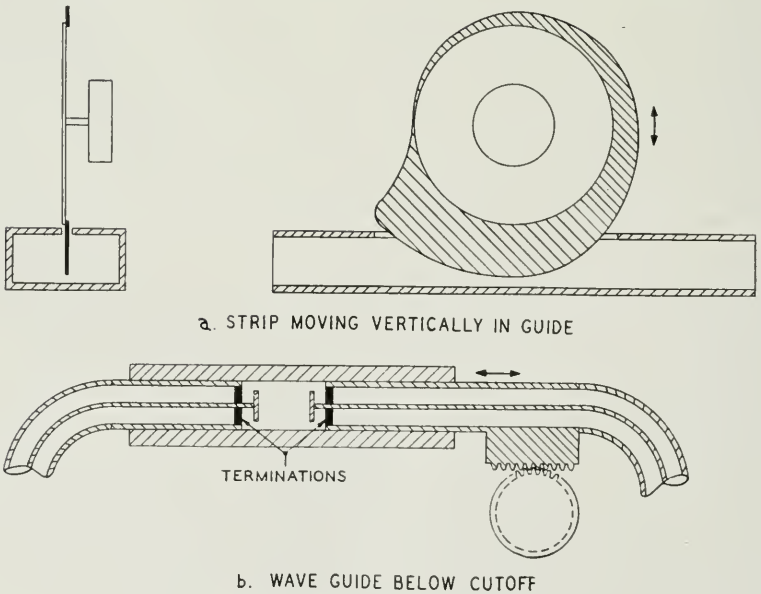
Fig. 23—Microwave attenuators.

ment and to obtain the minute test signals required for measuring receiver characteristics. Pads serve to change levels and to prevent interaction

between testing components. Microwave attenuators and pads are of two general types (a) those which employ dissipative elements to absorb power and (b) non-dissipative devices which introduce propagation or coupling loss.

For the shorter microwaves the most convenient form of attenuator is of the dissipative type, employing a strip or vane of dielectric coated with a resistance material, as for example, carbon-coated bakelite. This is placed in rectangular wave guide with its plane paralleling the side of the guide. The attenuation is varied by varying the depth to which the vane is inserted in the guide (Fig. 23a) or by changing its position in the guide. A valuable feature of such attenuators is that the minimum loss can be made substantially zero. For good impedance match the strip must be tapered. By using two strips the over-all length of the attenuator can be reduced. Extremely satisfactory attenuators of this type covering a frequency range of 8 to 12%, with loss variable from 0 to 35 or 40 db, have been obtained in the frequency range 4,000 to 24,000 mc.

For the longer microwaves, where wave guides are inconveniently large, attenuators of the wave guide-below-cutoff type are very useful. These consist of a section of round wave guide whose diameter is small compared with wavelength and whose length is adjusted by telescoping (see Fig. 23b). The $TM_{0,1}$ mode has been found very satisfactory, and $TE_{1,1}$ has also been used. Connection is made to the attenuator by a coaxial circuit at each end, with disk excitation for the $TM_{0,1}$ mode and loop coupling for $TE_{1,1}$. The attenuation formulas are:[12]

$TM_{0,1}$.
$$A = \frac{41.8}{D} \sqrt{1 - \left(\frac{1.31D}{\lambda}\right)^2} \; db/\text{meter} \tag{12}$$

$TE_{1,1}$.
$$A = \frac{32.0}{D} \sqrt{1 - \left(\frac{1.71D}{\lambda}\right)^2} \; db/\text{meter} \tag{13}$$

where D = diameter of wave guide in meters. Because of the effect of other modes when the coupling is close, a minimum loss of 20 to 30 db is required before the attenuation becomes linear with displacement. The attenuation differentials are substantially independent of frequency. Attenuators of this type present a large impedance mismatch at either end, the effect of which may be alleviated by padding or by a termination.

Types of pads employed in microwave work include the following:
(1) Flexible coaxial cable, usually with high resistance inner conductor.
(2) Coaxial π with carbon coated rod and discs.
(3) Coaxial with carbon coated rod as inner conductor.
(4) Resistance strip in wave guide.
(5) Directional coupler.

In calibrating microwave attenuators and pads, comparison with an accurately calibrated IF attenuator, using a heterodyne test set, has been found to give excellent results.

RF Cables and Connectors

Flexible RF cables for connecting test equipment to equipment under test are an important adjunct of field testing. At frequencies of 10,000 mc and below, flexible coaxial cables of about .4″ over-all diameter with solid or stranded inner conductor, solid low-loss dielectric (polyethylene) and braided outer conductor have been used satisfactorily, although in the upper part of this range special measures have been necessary to prevent attenuation change due to flexure and aging. Over most of this range coaxial jack and plug connections have been found satisfactory but wave guide connectors are preferable at the upper end. In the range above 10,000 mc, coaxial
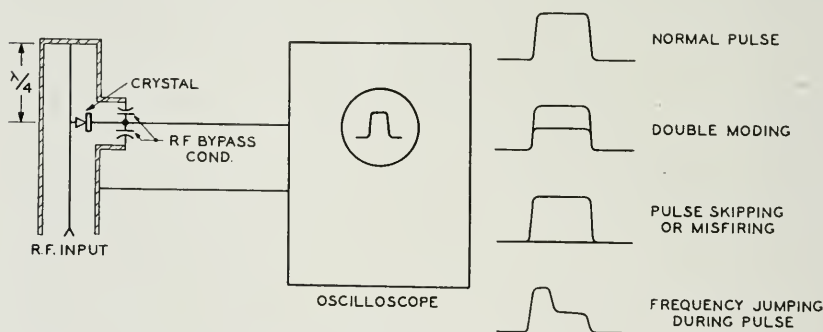
Fig. 24—Radar pulse envelopes.

cables of requisite stability have not yet been obtained and rubber covered wave guide with soldered articulated joints is the best form of flexible cable now available.

Oscilloscopes

Oscilloscopes are used extensively in radar maintenance (a) for examination of video waves and (b) for viewing RF envelopes. Satisfactory radar performance depends on a variety of video wave shapes which may include trapezoidal or triangular pulses, sawtooth waves, square waves or combinations of these. Observation of these wave shapes, supplemented if necessary by measurements of amplitude and duration, helps in diagnosing many troubles.

Examination of the envelope of the RF pulse is a convenient but less informative alternative to spectrum analysis. The envelope should be a clean, single trace of good shape. Figure 24 shows traces sometimes experi-

enced. Double moding, i.e., oscillating at different frequencies on different pulses, is shown by a double trace. Frequency jumping during a pulse is shown by a break in the envelope. Misfiring gives a base line under the envelope. Other abnormalities in the RF envelope may result from incorrect video wave shape. Observation of the RF envelope requires a rectifier, usually a crystal, together with a suitable video amplifier. Since limitation of the scope to video functions permits general application to radars of all frequencies, the rectifier is generally provided externally.

The oscilloscopes available before the war did not meet the requirements of radar. Fast sweeps were necessary to permit viewing of pulses ranging from several microseconds to a fraction of a microsecond. Amplifiers were required for such pulses with low phase and amplitude distortion over a broad frequency band. Existing methods of synchronizing and phasing sweeps were also inadequate. The progress of the oscilloscope art during the war is illustrated in the successive designs of field test oscilloscopes shown in Fig. 25.

The BC910A oscilloscope, gotten out as a "stop gap" not long after the attack on Pearl Harbor, incorporates fast sweeps and broad-band amplification. Following close upon this was the BC1087A (Navy code CW60AAY) which replaced sine wave synchronization by a start-stop sweep triggered by the incoming pulses. This feature made it possible to superpose the erratic pulses produced by spark wheel and similar pulsers and at the same time avoided external synchronizing connections. A valuable feature conjoined with the start-stop sweep was a delay network in the main transmission path which gave the sweep time to start before the pulse reached the cathode-ray tube. This oscilloscope in original and modified form has seen wide service in all theaters. However, its weight of more than 60 pounds was a handicap for many uses.

Further advances in oscilloscope circuitry and in weight limitation resulted in TS-34/AP, weighing only 25 pounds. This combined the short pulse features of the previous design with those of the conventional oscilloscope for viewing slower waves. A schematic diagram is shown in Fig. 26. A redesign, coded as TS-34A/AP, incorporated variable start-stop sweeps and improved mechanical design. These two oscilloscopes, TS-34 and TS-34A, were produced to a total of some 12,000 and universally used by all branches of the service for both radar and radio testing. Toward the end of the war the trend toward shorter pulses, coupled with the need for precise measurement of wave amplitude and duration, led to a new design, TS-239/UP, which embodied wide advances over TS-34A in performance and versatility but with an increase in weight.

In association with different oscilloscopes, other video devices have been
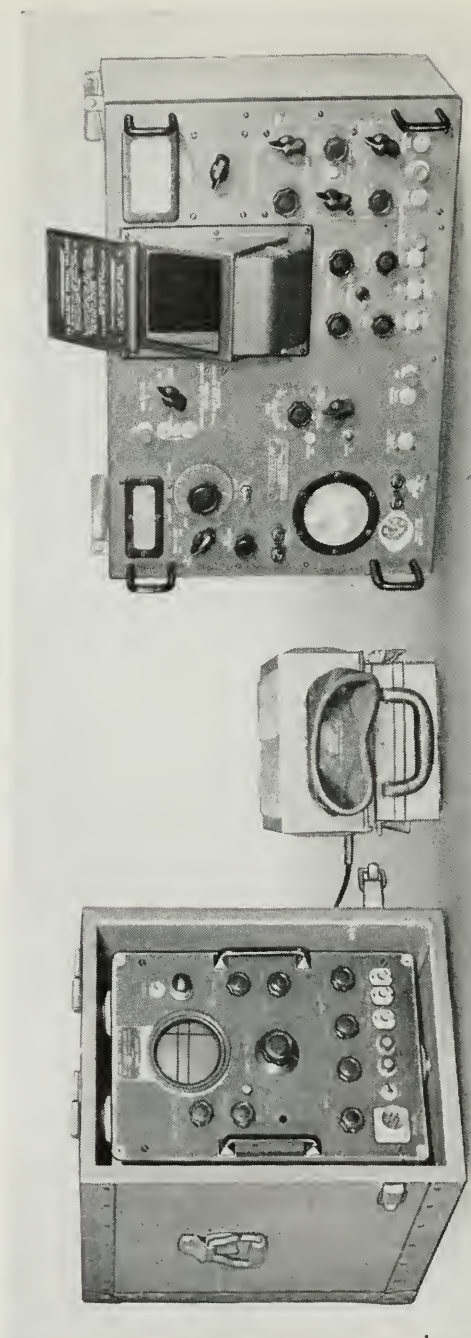
Fig. 25—Test oscilloscopes for viewing wave forms in radars. Left to right: BC-910-A (1942), TS-34A/AP (1944) and TS-239/Up (1945).

478

employed. The amplitude of pulse applied to the magnetron is thousands of volts. To derive a voltage suitable for application to the oscilloscope, a voltage divider of the condenser type is used (TS-89/AP). Suitable video terminations, dividers, and loads, sometimes of high voltage and power capacity, are required to obtain proper test conditions and provide convenient test points (TS-98/AP, TS-390/TPM-4, TS-90/AP, TS-234/UP). Originally a high-impedance connection to the oscilloscope was effected by a single-stage amplifier unit (BC1167A), but a simple divider type of probe was later found more satisfactory for this purpose.
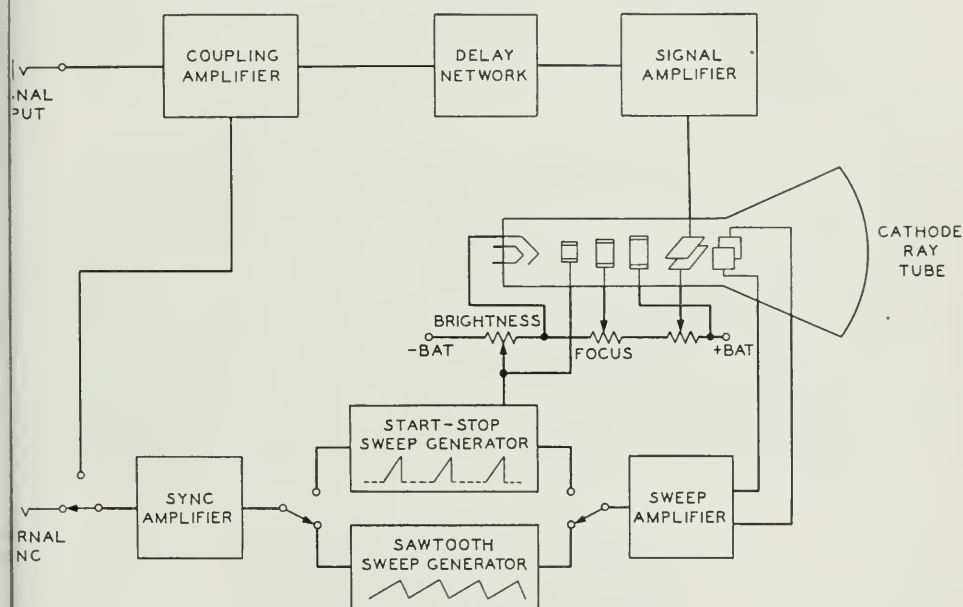
Fig. 26—Block diagram of TS-34/AP and TS-34A/AP oscilloscopes.

RANGE CALIBRATION

Types of timing circuits used for radar range determination include (a) multi-vibrators, (b) coil and condenser oscillators (generally without but sometimes with temperature control) and occasionally (c) quartz crystal oscillators. The first two depend for their accuracy on condensers, resistances, coils and other elements which are subject to error due to aging, temperature, humidity, mechanical damage and the like. Nor is the quartz crystal oscillator wholly immune to error. Consequently, portable range calibrators are required for field maintenance.

TS-102A/AP (Fig. 27) and its predecessors TS-102 and TS-19 are precision calibrators which have been extensively used for checking a large
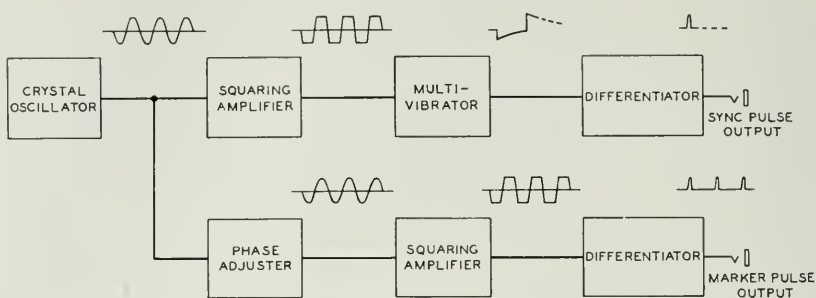
Fig. 27—Block diagram of TS-102/AP and TS-102A/AP range calibrators.
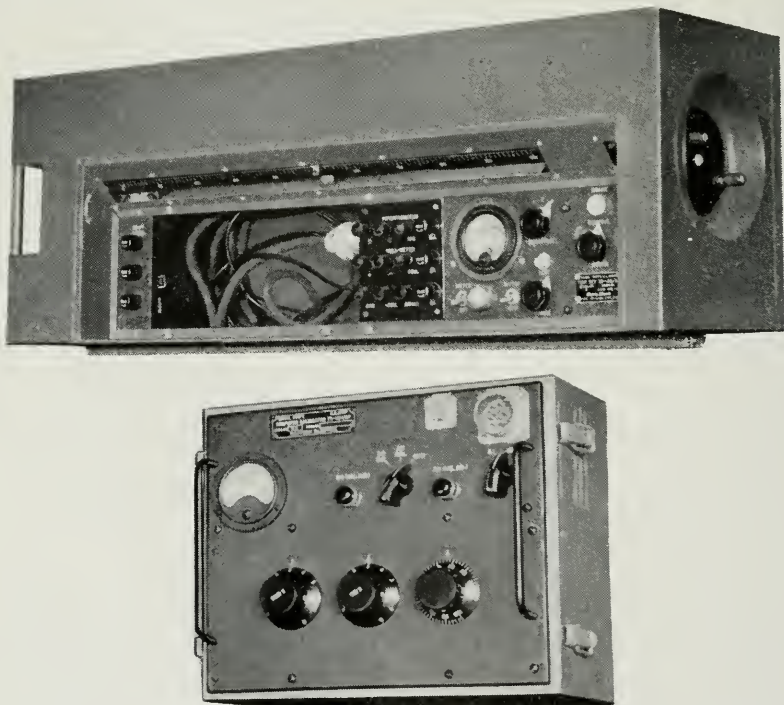
Fig. 28—Two test sets for checking the computers used in bombsight and fire control radars.

number of different airborne bombing and gunlaying radars and shipborne fire control and search radars. These sets deliver accurately spaced marker

pulses, derived from a quartz crystal oscillator, for checking the radar range pulses. A trigger pulse derived from a multi-vibrator synchronized with the quartz oscillator is also provided for actuating the radar timing circuits. With certain radars the calibration procedure requires an oscilloscope as well. Extreme stability of marker pulses, better than $\pm.02$ microsecond, is obtained. A stop watch is included in these sets for checking rate of change in range.

Less precision is required in range calibration of search radars. For this purpose the TS-5/AP calibrator provides marker pulses of $\frac{1}{4}$, 1, 5 or 10 nautical or statute miles, derived from a coil and condenser oscillator with closely controlled temperature coefficients. This calibrator is designed to be triggered by the radar or some other external source.

Computer Test Sets, Etc.

A number of radars are equipped with computers which receive the data on location of target and its direction and rate of change, together with essential related information on such factors as wind velocity, ground speed, altitude, etc., and deliver the solution of the ballistic problem in the form of a voltage which releases bombs, points the guns or serves other purposes. Means for checking the accuracy of these computing devices are generally required. The type of test set needed depends upon the computer design, which has taken different forms according to the nature of the problem and the state of the art.

Two types of computer test set are shown in Fig. 28. TS-158/AP, designed for use with certain airborne bombing radars, furnishes to the computer a signal representing a target approaching at known speed and checks the accuracy of bomb release. TS-434/UP, designed for several airborne and ground radars, is an accurate instrument for determining the voltage ratios at various points in a computer and thus checking its performance.

Conclusion

More than 200 different designs of test sets were developed during the war by Bell Laboratories to meet the exacting requirements of radar field maintenance. These differed radically from previous art. Outstanding features were portability, precision and generality of application. The large number of designs is due partly to the varied functions of radar and to the varied conditions of use. Largely, however, it results from the fact that the frequency band that can be handled in any one set is limited, whereas many frequency ranges and subranges had to be covered in all.

Altogether more than 75,000 radar test sets were manufactured by Western Electric Company and these were used in all theatres of war by the

United Nations forces. The production rate at the end of the war exceeded 5,000 test sets a month. In numerous cases, moreover, small preproduction quantities of test equipment were built on a "crash" basis for special missions and for training purposes. The test equipment produced for the field had to be more precise than the radars, and the equipment used in the factory and laboratory to test the field test equipment had to be still more precise.

Trends of development at war's end were toward (a) further broad-banding, simplification and precising, and (b) coverage of new frequency ranges.

REFERENCES

(1) "Ultra-High Frequency Techniques," by J. G. Brainard, G. Koehler, H. J. Reich, and L. F. Woodruff, D. Van Nostrand, 1942.
(2) "Reflex Oscillators," by J. R. Pierce, *Proc. I.R.E.*, Vol. 33, Feb. 1945, p. 112.
(3) "The Lighthouse Tube," by E. D. McArthur and E. F. Peterson, *Proceedings of National Electronics Conference*, Vol. I, 1944, p. 38.
(4) "Electromagnetic Waves," by S. A. Schelkunoff, D. Van Nostrand, 1943.
(5) "The Proportioning of Shielded Circuits," by E. I. Green, F. A. Leibe and H. E. Curtis, *Bell System Technical Journal*, Vol. 15, April 1936, p. 248.
(6) "Resonant Lines in Radio Circuits," by F. E. Terman, *Electrical Engineering*, Vol. 53, July 1934, p. 1046.
(7) "Ultra-Short-Wave Transmission Phenomena," by C. R. Englund, H. T. Crawford, and W. W. Mumford, *Bell System Technical Journal*, Vol. 14, July 1935, p. 369.
(8) "Thermistors in Electronic Circuits," by R. R. Batcher, *Electronic Industries*, Vol. 4, Jan. 1945, p. 76.
(9) "An Impedance Transmission Line Calculator," by P. H. Smith, *Electronics*, Jan. 1944.
(10) "Microwave Transmission," by J. C. Slater, McGraw-Hill, 1942.
(11) "Theory of the Electromagnetic Horn," by W. L. Barrow and L. J. Chu, *Proc. I.R.E.*, Vol. 27, Jan. 1939, p. 51.
(12) "Attenuation of Electromagnetic Fields in Pipes Smaller Than Critical Size," by E. G. Lindner, *Proc. I.R.E.*, Vol. 30, Dec. 1942, p. 554.

Performance Characteristics of Various Carrier Telegraph Methods

By T. A. JONES and K. W. PFLEGER

This paper describes laboratory tests of certain carrier telegraph methods, to determine their relative advantages from the standpoints of signal speed, and sensitivity to level change, carrier frequency drift, interchannel interference, and line noise.

Introduction

MOST of the carrier telegraph methods mentioned below are well known,[1,2,3,4,5] but the selection of the best method for a particular application is difficult without comparative tests on specific designs. It is the purpose of this paper to record data taken during such tests and to explain the results so that they may be helpful to those concerned with the selection of the optimum method for a given set of requirements.

The conclusions here reached regarding methods of telegraph transmission do not necessarily apply to transmission of sound, pictures, or television, because their requirements differ. In telegraph transmission it is important that signal transitions be received at approximately the correct times, and wave rounding is permissible.

Computations for a square cut-off band-pass filter with zero phase distortion[6] show that the shape and duration of the transient in the received wave are about the same for a sudden transition in both on-off and frequency-shift arrangements (explained in the next section), when the total frequency shift is not more than half the channel width. As telegraph distortion depends largely upon the transient, one might therefore infer that, if the transients are about alike, there is no particular advantage in frequency-shift over the on-off method as far as signal speed is concerned. However, the computation for the idealized filter gives no assurance that a physical filter will per-

[1] H. Nyquist: "Certain Topics in Telegraph Transmission Theory", *A. I. E. E. Trans.*, Vol. 47, pp. 617–644, April 1928.

[2] H. Nyquist and K. W. Pfleger: "Effect of Quadrature Component in Single Sideband Transmission", *Bell System Technical Journal*, Vol. XIX, pp. 63–73, Jan. 1940.

[3] E. H. Armstrong: "Methods of Reducing the Effect of Atmospheric Disturbances", *Proc. I. R. E.*, Jan. 1928, pp. 15–26.

[4] J. R. Carson: "Reduction of Atmospheric Disturbances", *Proc. I. R. E.*, July 1928, pp. 966–975.

[5] F. B. Bramhall & J. E. Boughtwood: "Frequency Modulated Carrier Telegraph System", *Electrical Engineering*, Vol. 61, No. 1, Jan. 1942, *Transactions Section*, pp. 36–39.

[6] Fig. 3 of H. Salinger: "Transients in Frequency Modulation", *Proc. I. R. E.*, August 1942, pp. 378–383.

form thus. In order to investigate this experimentally, as well as other factors that concern the choice of method, the effects on telegraph transmission of interchannel interference and of varying the signaling speed, transmission level, mean carrier frequency, and line noise, were determined for several different methods, using the same channel filters. In order to test the two-band methods* using the same frequency range occupied by the one-band arrangements, narrow-band filters would be required to divide the frequency range into two parts. Since such filters were not available, it was necessary to use two adjacent frequency bands each similar to that used with the on-off method. However, some tests were made of a two-band arrangement using somewhat narrower filter pass bands.

A special wide-band frequency-shift arrangement using filters of about twice the band width of the other frequency-shift arrangement, was tested mainly in order to observe the effect of band width on sensitivity to noise and interference.

In all of the noise tests, thermal or resistance noise was used. With noise of the impulse type it is possible that somewhat different results would have been obtained, but it is believed that the difference would not have been great.

Conclusions

A study of the test results leads to the following conclusions which apply for the conditions assumed, and which are thought to be of general application, except for modifications which may be made necessary by future technical advances:

1. There is no important advantage in frequency-shift carrier telegraph over the on-off method as used in the Bell System for stable, quiet circuits, either wire or radio. However, the frequency-shift method shows some improvement in operating through noise. The frequency-shift method has disadvantages as regards complication and cost. Furthermore, it may be seriously affected by carrier frequency drift and interchannel interference, although the effects of these can be mitigated to some extent by special devices.

2. For high-frequency radio transmission over long distances, which is subject to comparatively severe non-selective fading, a great advantage is realized from the use of frequency-shift telegraphy with a fast receiving limiter instead of the conventional "continuous wave" or on-off method. For satisfactory operation it is still necessary that the signal level be kept sufficiently higher than the noise level in the transmission band.

* See the section entitled "Explanation of Terms".

3. Single-sideband telegraphy has an advantage of providing somewhat higher speeds without increasing the band width. Whether it holds much promise for any general application in multi-channel systems utilizing narrow bands and moderate signal speeds is questionable in view of certain difficulties. For a single-channel high-speed circuit, single-sideband telegraphy might be found worth while from the standpoint of economical use of the frequency spectrum.

4. Certain other arrangements tested possessed some characteristics which have advantage under particular conditions. For example, two-band arrangements may sometimes be conveniently obtained by combining existing on-off arrangements. These two-band arrangements are capable of furnishing high-grade service over radio circuits subject to severe fading. The use of a single source of carrier instead of two sources on a two-band arrangement results in a substantial transmission improvement. The performance then is comparable to that of a single-band frequency-shift channel occupying the same frequency space.

A more complete discussion of the results is given under the heading "Summary of Results", at the end of this paper.

Explanation of Terms

The following is intended to explain what is meant by certain terms used in this paper. They apply specifically to carrier telegraph operation in the voice range but, in general, they could also apply to radio telegraphy. (It will be appreciated that various other combinations of the instrumentalities involved in the present discussion could be used.)

Channel

A telegraph channel is a path which is suitable for the transmission of telegraph signals between two telegraph stations. In the present discussion the term "channel" is restricted to mean one of a number of paths for simultaneous transmission in different frequency ranges as in carrier telegraphy, each channel consisting of an arrangement of carrier telegraph equipment designed for the transmission of one message at a time, in only one direction.

On-Off Method

This, the most common form of amplitude modulation, is the same as "continuous wave" in radio telegraphy. It is a method of signaling over a channel utilizing a single carrier frequency, normally located at the center of the transmission band of the channel filters. The presence of carrier current on the line corresponds to the marking condition of the channel, and its absence, to the spacing condition. A Fourier analysis of the line current

during signaling would show a steady carrier frequency component and substantially symmetrical upper and lower sideband frequency components.

Single-Sideband Method

This is similar to the method just described except that: (1) the carrier frequency is located near one boundary of the channel filters, so that during signal transmission one of the sidebands is attenuated much more than the other before the signals reach the line, and (2) during the spacing condition, carrier current may be either absent from the line, or present with amplitude less than that of marking current. (The latter condition tends to reduce that part of the distortion which is due to the quadrature component.[2])

Frequency-Shift Method

This is a method of signaling over a channel utilizing a carrier current of substantially constant amplitude from a frequency-modulated oscillator. The carrier current has no phase discontinuity and its instantaneous frequency varies between two limits within the transmission band. In the present discussion the two limits, symmetrically located in the transmission band, correspond respectively to the marking and spacing conditions of the channel. Variation of the instantaneous frequency may be abrupt or gradual, for example, sinusoidal. Throughout this paper the reader should assume that the variation is substantially abrupt except where otherwise indicated. At the receiving terminal the variable frequency signals are converted to amplitude-modulated signals by means of a frequency detector.

Two-Source Method

This is also a frequency-shift method, but it differs from that described above in that it is made up by combining two on-off channels, each supplied with carrier current of a different but substantially constant frequency from a separate source. In the present discussion each of the two on-off channels has a separate oscillator and occupies a different frequency band on the same line, and the sending relays of the channels have their operating windings differentially inter-connected so that one oscillator delivers current to the line during marks, and the other during spaces. Thus the marking and spacing signals are confined to separate frequency bands on the line. This is sometimes referred to as two-band operation. The switching takes place abruptly. No attempt is made to control the phases of the two sources. Therefore phase discontinuities are likely to occur at the instants of switching, causing brief transients of varying shapes in the line current. The output circuits of the two receiving detectors are differentially interconnected to obtain polar signals for the operation of a common receiving relay.

One-Source Two-Band Method

This is also a frequency-shift method, and is substantially the same as the two-source method except that a single frequency-modulated oscillator is used at the sending end instead of two oscillators, and thus there is no phase discontinuity in the sent signals.
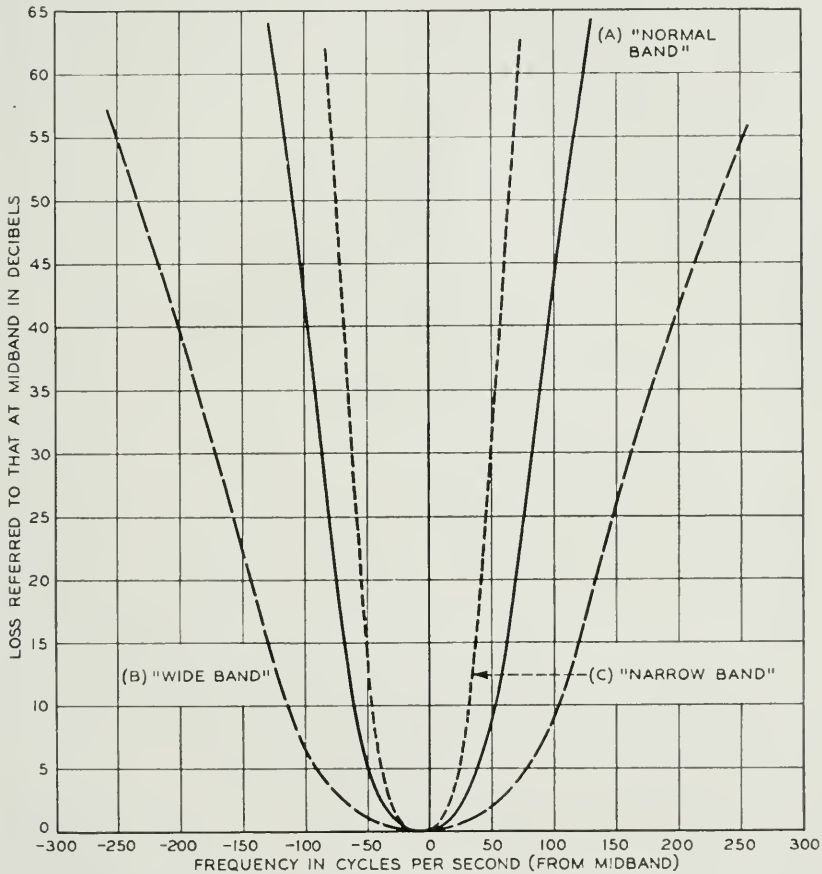
Fig. 1—Loss characteristics of channels tested, including sending and receiving filters and two repeating coils.

These methods will be more clearly understood from the following description of apparatus.

Apparatus

Channel Filter Characteristics

In Fig. 1, curve A represents the loss vs. frequency characteristic of the majority of channels used in the tests, including both sending and receiving

filters and associated repeating coils. The two-source arrangements each occupied two such bands with midband frequencies spaced 170 cycles apart, except in the case of the narrow-band two-source arrangement which occupied two bands having characteristics similar to curve C with midband frequencies spaced 120 cycles apart. Curve B represents the loss vs. frequency characteristic of the wide-band frequency-shift arrangement, having approximately twice the band width of curve A. These characteristics were all measured between 600-ohm terminations, without adjacent channel filters present.

On-Off Terminal Apparatus

Figure 2 shows in block form the different circuit arrangements tested, together with the location of the carrier frequencies in the transmitted bands. In the on-off arrangement, the oscillator transmitted 1955-cycle current to a modulator which contained a polar telegraph relay controlled by signals from the local sending loop. During spacing signals this relay short-circuited the carrier supply, and during marking signals it allowed the carrier current to flow through the sending band-pass filter to the adjustable resistance line.

When it was desired to measure the effect of interference from other channels working in adjacent pass bands, their terminal equipment was added by connecting the line sides of their sending or receiving filters to the common sending or receiving bus (See dashed lines designated "bus" in Fig. 2), so that all the channels would transmit over the same line.

After the carrier signals passed through the receiving filter connected to the output of the line, they were converted by a detector-amplifier into direct current for operating a polar receiving relay, which, in the absence of incoming signals, was held on its spacing contact by local biasing current. The receiving relay contacts transmitted into the local receiving loop.

Single-Sideband Terminal Apparatus

In testing the single-sideband arrangement shown in Fig. 2, the terminal equipment of the on-off arrangement was used and the carrier frequency was placed 43 cycles above midband, at 1998 cycles. In some of the tests the modulator was modified to transmit during spacing intervals a carrier current 6 db below its marking value. Since the loss in the receiving filter was greater at the edge of the band than at the center, it was necessary to increase the gain of the linear detector-amplifier in order to have the same change of current in the line winding of the receiving relay of the single sideband arrangement as in the on-off arrangement. A further increase in the detector-amplifier gain was necessary when the single-sideband arrangement was operated with spacing carrier 6 db below the marking carrier.

ARRANG...

ON - ...

D-C
AMPLIFIER

RECEIVING
POLAR
RELAY

BIAS
WINDING

RECEIVING
LOOP

SIN...
SIDE...

FREQU...
SH...

MARKING
DETECTOR

MARKING
LOW-PASS
FILTER

SPACING
DETECTOR

SPACING
LOW-PASS
FILTER

PUSH-PULL
D-C
AMPLIFIER

RECEIVING
POLAR
RELAY

RECEIVING
LOOP

TW...
SOU...

MARKING
AMPLIFIER-
DETECTOR

MARKING
LOW-PASS
FILTER

SPACING
AMPLIFIER-
DETECTOR

SPACING
LOW-PASS
FILTER

PUSH-PULL
D-C
AMPLIFIER

RECEIVING
POLAR
RELAY

RECEIVING
LOOP

ONE-S...
TWO-...

-SOURCE ARRANGEMENT
USED)

Fig. 2—Telegraph arrangements tested.

Frequency-Shift Terminal Apparatus

In most of the tests on the frequency-shift arrangement shown in Fig. 2, the oscillator frequency was caused to vary abruptly by a relay modulator. The sending relay was of the same type as used in the on-off arrangement and varied the tuning capacity of the oscillator. For some other tests the frequency variation was made more gradual by converting the sent signals into polar signals and passing these through a low-pass filter in order to round the wave so that the pulses had an approximately sinusoidal shape during reversals and attained steady-state value only at the center of each pulse. These polar signals were used to control the plate resistance of either of two diodes, thereby connecting a positive or negative reactance across the tuned circuit of the oscillator.[5] This caused the oscillator frequency to be either increased or decreased in proportion to the amplitude change of the control current. During most of the tests the spacing frequency was 1920 cycles and the marking frequency was 1990 cycles, both equally spaced from the midband frequency, 1955 cycles. The oscillator, when on marking frequency, was set to produce one milliwatt into the 600-ohm resistance artificial line. After passing through the receiving filter connected to the output of the artificial line, the signals entered a limiter (unless otherwise stated) delivering an output current which was practically constant for input levels between -55 and $+25$ dbm. From the limiter the signal passed into a frequency discriminator circuit having two output branches, each of which was connected to a diode detector tube followed by a low-pass filter. The two discriminator branch circuits in combination with their detectors and low-pass filters had output amplitude vs. input frequency characteristics of opposite slopes. After differential recombination of the two low-pass filter outputs, the resultant characteristic was linear over the range of fundamental frequencies transmitted by the limiter. The differentially recombined wave in the final d-c amplifier had an amplitude substantially proportional to the instantaneous deviation from the average value of the received carrier frequency over a range of ± 70 cycles. (Some calculations by one of the writers indicate that the use of discriminators of this type is helpful in reducing characteristic telegraph distortion.) In most of the tests the low-pass filters associated with the detector output had a cut-off frequency (about 503 cycles) low enough to suppress the carrier but high enough not to affect the telegraph transmission. The final d-c amplifier was substantially linear and increased the d-c wave to a suitable value for operating the polar receiving relay.

The wide-band frequency-shift arrangement was similar to that just described, except for the change in filters and tuning of the sending oscillator and the discriminator. The spacing frequency was 2055 cycles and the

Frequency-Shift Terminal Apparatus

In most of the tests on the frequency-shift arrangement shown in Fig. 2, the oscillator frequency was caused to vary abruptly by a relay modulator. The sending relay was of the same type as used in the on-off arrangement and varied the tuning capacity of the oscillator. For some other tests the frequency variation was made more gradual by converting the sent signals into polar signals and passing these through a low-pass filter in order to round the wave so that the pulses had an approximately sinusoidal shape durin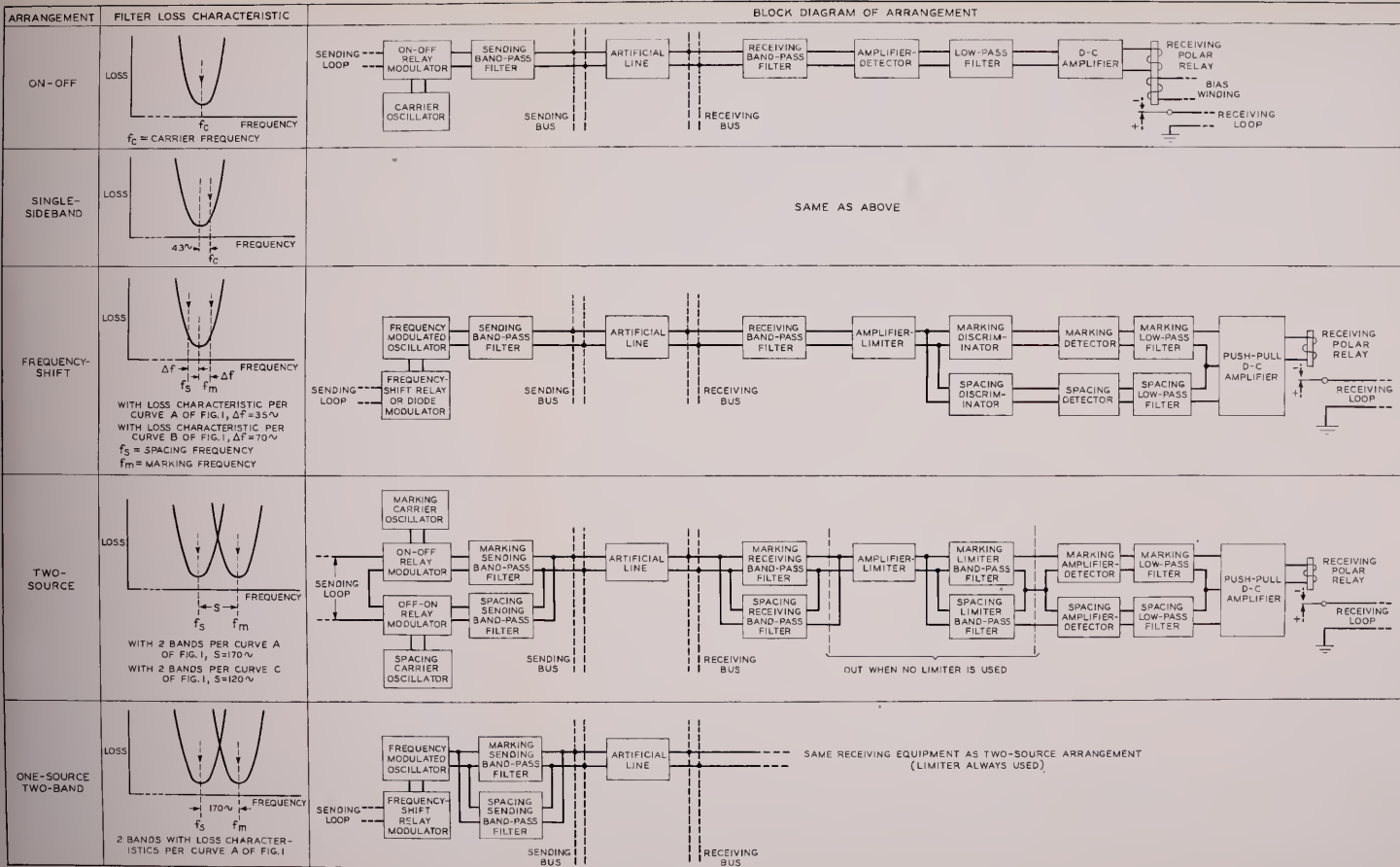g reversals and attained steady-state value only at the center of each pulse. These polar signals were used to control the plate resistance of either of two diodes, thereby connecting a positive or negative reactance across the tuned circuit of the oscillator.[5] This caused the oscillator frequency to be either increased or decreased in proportion to the amplitude change of the control current. During most of the tests the spacing frequency was 1920 cycles and the marking frequency was 1990 cycles, both equally spaced from the midband frequency, 1955 cycles. The oscillator, when on marking frequency, was set to produce one milliwatt into the 600-ohm resistance artificial line. After passing through the receiving filter connected to the output of the artificial line, the signals entered a limiter (unless otherwise stated) delivering an output current which was practically constant for input levels between -55 and $+25$ dbm. From the limiter the signal passed into a frequency discriminator circuit having two output branches, each of which was connected to a diode detector tube followed by a low-pass filter. The two discriminator branch circuits in combination with their detectors and low-pass filters had output amplitude vs. input frequency characteristics of opposite slopes. After differential recombination of the two low-pass filter outputs, the resultant characteristic was linear over the range of fundamental frequencies transmitted by the limiter. The differentially recombined wave in the final d-c amplifier had an amplitude substantially proportional to the instantaneous deviation from the average value of the received carrier frequency over a range of $\pm$ 70 cycles. (Some calculations by one of the writers indicate that the use of discriminators of this type is helpful in reducing characteristic telegraph distortion.) In most of the tests the low-pass filters associated with the detector output had a cut-off frequency (about 503 cycles) low enough to suppress the carrier but high enough not to affect the telegraph transmission. The final d-c amplifier was substantially linear and increased the d-c wave to a suitable value for operating the polar receiving relay.

The wide-band frequency-shift arrangement was similar to that just described, except for the change in filters and tuning of the sending oscillator and the discriminator. The spacing frequency was 2055 cycles and the

marking frequency was 2195 cycles which were equally spaced from the midband frequency, 2125 cycles. In this arrangement a limiter was always used at the receiving terminal. The discriminator was adjusted to be linear over twice the frequency range of the previously described discriminator, and the slope of the new discriminator characteristic was adjusted to give the same marking or spacing output as before. The wide-band frequency-shift arrangement was also tested with low-pass filters associated with the detector outputs which had a cut-off frequency (about 58 cycles) low enough to give a distortion vs. speed characteristic close to that obtained on the normal-band frequency-shift arrangement. The loss of each of these low-pass filters was about 10 db at 58 cycles. Since each filter consisted of only one section the cut-off was gradual.

Two-Source Terminal Apparatus

The sending circuits of the two-source arrangement shown in Fig. 2 included separate oscillators of different frequency for marking and spacing signals. Two sending relays were operated in synchronism by the signals in the sending loop. The relay contacts were so connected that marking carrier was transmitted to a marking band-pass filter and spacing carrier was cut off from a spacing band-pass filter, or vice versa. Thus either the marking or the spacing carrier frequency was transmitted to the line at any instant. At the receiving end of the line the incoming signals flowed through two receiving filters, one passing the marking current and the other passing the spacing current. When a limiter was not used, the outputs of these filters were connected directly to separate amplifiers, detectors and low-pass filters. The rectified marking and spacing signals were combined differentially, passed through a push-pull d-c amplifier, and operated the receiving relay just as in the frequency-shift arrangement. When a limiter was used the outputs of the receiving filters were recombined and passed through the limiter, after which they were again separated by means of additional band-pass filters whose losses were about half those of the receiving band-pass filters. The two-source arrangement was tested both with and without limiter when the loss characteristics of the spacing and marking paths were each similar to curve A of Fig. 1 and had midband frequencies of 1785 and 1955 cycles, respectively.

The two-source arrangement with limiter also was tested with filters having loss characteristics for the spacing and marking paths each similar to curve C of Fig. 1, and having midband frequencies of 1980 and 2100 cycles, respectively. This arrangement occupied a frequency band approximately 12/17 that used for the two-source arrangement with filters having characteristics similar to curve A.

One-Source Two-Band Terminal Apparatus

The one-source two-band sending circuit was similar to that of the frequency-shift arrangement with relay modulator, except that the frequency shift was 170 cycles, and the marking and spacing frequencies were adjusted to be at the centers of the pass bands of the marking and spacing sending filters, as shown in Fig. 2. The receiving circuit was the same as that used for the two-source arrangement. A limiter was always used.

Telegraph Transmission Measuring Apparatus

In all of the tests, the d-c open-and-close signals in the local sending loop were substantially rectangular and consisted either of reversals (a succession of alternate marks and spaces of equal duration) or of the test sentence: THE QUICK BROWN FOX JUMPED OVER A LAZY DOG'S BACK 1234567890 BTL SENDING. The customary 7.42 unit teletypewriter code was used, consisting of a stop pulse, a start pulse, and five code pulses per character[7]. The distributors supplying these signals were driven by synchronous motors controlled by an adjustable frequency oscillator. The speeds utilized experimentally ranged from 60 to 180 words per minute (about 23 to 68 dots per second).

In order to measure the telegraph distortion of the signals obtained in the receiving loop, a cathode-ray tube distortion measuring set was used which measured maximum total distortion in per cent of a unit pulse in much the same manner as a start-stop distortion measuring set previously described[8], except that electronic circuits were used to replace the distributor and all but one relay, which made possible precise measurements over a wide range of speeds. The bias of received signals was measured on reversals by means of a highly damped zero-center d-c milliammeter inserted in the receiving loop.

Source of Resistance Noise

In order to measure the effect of line noise, resistance noise was reproduced from a phonograph record, amplified, and combined with the carrier signals by means of a symmetrical three-way pad (part of the artificial line). A variable attenuator was used to regulate the amount of noise entering the line. The r.m.s. noise power or marking carrier power was measured with a thermocouple.

Measuring Precision

The signals generated by the dot and test sentence distributors were distorted less than 3 per cent of a dot length. As these distributors were of a

[7] E. F. Watson: "Fundamentals of Teletypewriters Used in the Bell System", *Bell Sys. Tech. Jour.*, Vol. XVII, Oct. 1938, pp. 620–639.

[8] R. B. Shanck, F. A. Cowan, S. I. Cory: "Recent Developments in the Measurement of Telegraph Transmission", *Bell Sys. Tech. Jour.*, Vol. XVIII, Jan. 1939, p. 149.

commercial type, they are believed to be representative. This distortion was erratic, depending upon the speed and wear of brushes and commutators. The distortion measured at the receiving relay could not be corrected by subtracting the distortion of the sent signals because during miscellaneous signals maximum distortion in the received signals might occur on a different transition from that in the sent signals. The small errors which existed in the sent signals are therefore believed to have been neither serious in their effect on the measured distortion, nor the sole cause of irregularity in the data.

The accuracy of the distortion measuring set itself was in the order of 1 per cent, as determined by measuring known amounts of bias in signals sent from a special distributor. Usually two observers took independent readings which were required to check closely or the observations were repeated. The average of the two observations was taken as the final measurement.

Although the line loss was constant in these tests, amplifiers, oscillators, power packs, and telegraph batteries were subject to slight voltage variations. Precautions were taken to reduce all variables as far as was practicable, yet it seems likely that the telegraph transmission measurements may be slightly in error due to such variations.

The individual sources of error mentioned in the last three paragraphs seem reasonable and sufficient to account for most of the irregularities in the following curves of telegraph transmission vs. speed. Yet it did not seem fair to draw smooth curves and neglect the irregularities, because these can also be due to the telegraph system itself, as was found by careful and repeated measurements. For example, it is known that relay performance is erratic and depends upon the speed. Chattering of relay contacts and periodic vibration of the armature have appreciable effect upon the distortion and can cause irregularities in distortion vs. speed characteristics, particularly at the higher speeds. Furthermore, such irregularities may not be wholly reproducible in repeated measurements due to changes in the relay temperature or contact surfaces and due to the occasional readjustments of relays. Another cause of irregularities in a distortion vs. speed characteristic may be the loss and phase characteristics of the channel filters. For example, consider an ideal transducer[1] which is distortionless at a speed s near the cut-off. It is also distortionless at speeds such as s/2, s/3, s/4, s/5, etc. At intermediate speeds distortion may exist, so that a curve of distortion vs. speed would show irregularities with minima at these optimum speeds. (Some computations by one of the writers for a frequency-shift arrangement using idealized filters show irregularities in the distortion vs. duration curve for a single dot.) No attempt was made to shape the channel filter characteristics to be perfectly distortionless at a particular speed; and it seems

reasonable that there could likewise be a number of speeds where maxima and minima occur in the distortion.

It is practically impossible to sift out and measure all causes of irregularities in a reasonable time. Therefore the irregularities have been shown exactly as measured in all the following curves wherein noise and interchannel interference were absent.

Transmission fluctuations due to the causes just mentioned were small compared to those encountered when strong interference or noise was present, because the latter varied greatly with time. In resistance noise, for example, peaks of great amplitude occur occasionally, although most of the time the fluctuations are relatively minor. It was necessary to observe for several minutes the distortion measured in the presence of resistance noise before one could be sure of finding anything approaching the maximum distortion; and the longer the period of observation, the greater was the peak distortion. In order to complete the testing in a reasonable time, watching periods were restricted to five minutes per observer and his maximum distortion reading was recorded. The results of two such observation periods for the same noise condition were averaged to determine a point for an experimental curve of distortion vs. noise-to-carrier ratio. Such points when plotted failed to lie in a perfectly smooth curve, but a smooth curve disregarding irregularities was drawn through the available points in what was estimated to be the correct location. The curves are described under the heading "Noise Tests" and may be used for comparison purposes, but are not an exact measure of the worst distortion to be expected over a long period of time. A similar procedure was followed for distortion measurements with interchannel interference.

Distortion vs. Speed Tests

The distortion mentioned throughout this paper is the absolute value of the maximum total distortion measured with the test sentence, and for brevity is merely called distortion. Except where otherwise specified, the arrangements were previously adjusted to have zero bias on reversals at the same speed. In these tests line noise and interchannel interference were absent.

Frequency-Shift Arrangements

Limiter

Some preliminary measurements on a frequency-shift arrangement having the loss characteristic of curve A of Fig. 1, with carrier varied abruptly from 1920 to 1990 cycles, showed that the limiter has little effect on the speed of the channel whether or not channel filters are used. The fact that low distortion was measured without channel filters indicated that the channel and

measuring devices were in good condition as far as could be reasonably expected over the range of speeds. The distortion measured without channel filters ranged from 2.5 per cent at 60 w.p.m. (23 d.p.s.) up to 10 per cent at 170 w.p.m. (65 d.p.s.).

Swing

Some measurements were made on the complete frequency-shift arrangement over the same range of speeds, using several values of abrupt frequency swing from ± 15 cycles to ± 55 cycles, keeping the marking and spacing frequencies equidistant from 1955 cycles. With a swing of ± 55 cycles the distortion was slightly worse than when the swing was ± 35 cycles. The measurements showed least distortion for a swing of ± 15 cycles. It has been previously shown[6] that the less the swing the smaller the amplitude of the oscillations in the transient for a given channel frequency band width. Accordingly one might expect distortion to be least when the swing is least. If only a small swing is used the signal bias change with carrier frequency drift is worse, unless automatic bias compensation is provided. Greater amplification is also required in the detector-amplifier in order to maintain the same relay operating current. A swing of ± 35 cycles was used in most of the frequency-shift tests as a good compromise between the distortion caused by the greater swings and the severe apparatus requirements and greater susceptibility to noise when using the lesser swings.

Types of Modulator

Figure 3 shows distortion characteristics of a frequency-shift arrangement using different types of modulator. Curve A was measured with sinusoidal frequency variation obtained by the use of a diode modulator and low-pass filter at the modulator input[5]. When the low-pass filter was omitted, the frequency variation was substantially abrupt, and curve B was obtained. When the diode modulator was replaced by a relay modulator, which also produced a substantially abrupt frequency change, curve C resulted. There is not much difference between these characteristics at low speeds. At high speeds the abrupt frequency variation appears to give somewhat lower distortion than sinuosidal variation. The distortion shown by curve A depends not only upon the speed and channel filter characteristic but also upon the characteristic of the low-pass filter used in rounding the sent wave in order to produce sinusoidal frequency variation. A considerable amount of care was necessary to prevent this low-pass filter from introducing too much distortion and at the same time to produce sufficient rounding. The cut-off frequency of this low-pass filter was adjusted at each signaling speed to be about three times the dot frequency. It was apparently low enough to
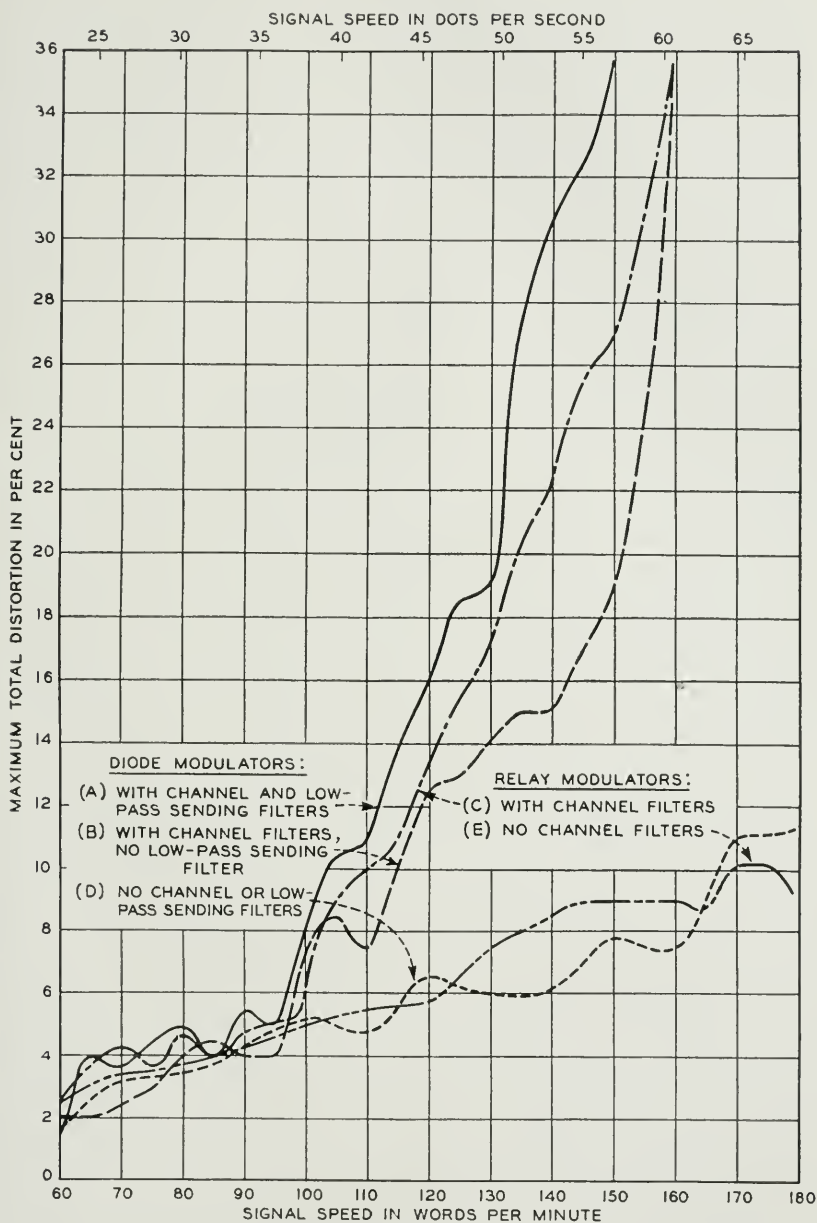
Fig. 3—Distortion vs. speed characteristics of normal-band frequency-shift arrangements, using two different types of modulator.

cause some characteristic distortion. No attempt was made to obtain an ideal distortionless filter characteristic because it would have been necessary also to take into account the characteristics of other parts of the circuit, which would have required considerably more time than was then available. Curves D and E of Fig. 3 show that the modulators caused very little distortion when channel and low-pass sending filters were absent.

Band Width

Figure 4 shows a comparison of the distortion vs. speed characteristics of the normal-band and wide-band frequency-shift arrangements when a relay modulator was used. Curve A of Fig. 4 is the same as curve C of Fig. 3 and shows the characteristic of the normal band arrangement. Curve C of Fig. 4 shows the characteristic of the wide-band arrangement when the low-pass filters at the detector output were the same as for the normal-band arrangement, their cut-off frequency being about 503 cycles. The distortion shown in curve C is much lower than that of curve A since the width of the side-bands transmitted was doubled. Curve B of Fig. 4 shows the characteristic of the wide-band arrangement when the low-pass filters at the detector output had a cut-off at about 58 cycles. As previously indicated, the latter cut-off was selected in order to give the wide-band arrangement about the same distortion vs. speed characteristic as the normal-band arrangement. The reason for the use of the lower cut-off is explained under the heading "Noise Tests".

Comparison of Frequency-Shift and On-Off Arrangements

Figure 5 is a comparison of the distortion vs. speed characteristic of the frequency-shift arrangement having a relay modulator (curve B), with characteristics of two on-off arrangements having commercial receiving circuits[9, 10, 11]. The 40B1 detector had no level compensator and included a triode detector having an output vs. input characteristic which roughly followed a square law. The other detector had a slow acting level compensator[11] designed to eliminate receiving bias due to slow changes in line equivalent. The output vs. input characteristic of this detector was much steeper than that of the 40B1 arrangement at the transition points of the signals. The level compensated arrangement was adjusted at each speed to

[9] B. P. Hamilton, H. Nyquist, M. B. Long and W. A. Phelps: "Voice Frequency Carrier Telegraph System for Cables", *Jour. A. I. E. E.*, Vol. XLIV, No. 3, Mar. 1925.

[10] A. L. Matte: "Advances in Carrier Telegraph Transmission", *B. S. T. J.*, Vol. XIX, pp. 161–208, Apr. 1940.

[11] A separate paper describing a commercial system using an improved level compensator is now in preparation by other Bell System authors. The function of the level compensator is to vary the gain of the receiving amplifier-detector so as to automatically compensate for relatively slow level changes. See also: V. P. Thorp: "A Level Compensator for Carrier Telegraph Systems", *Bell Laboratories Record*, Vol. XVIII, No. 2, October 1939, pp. 46–48.
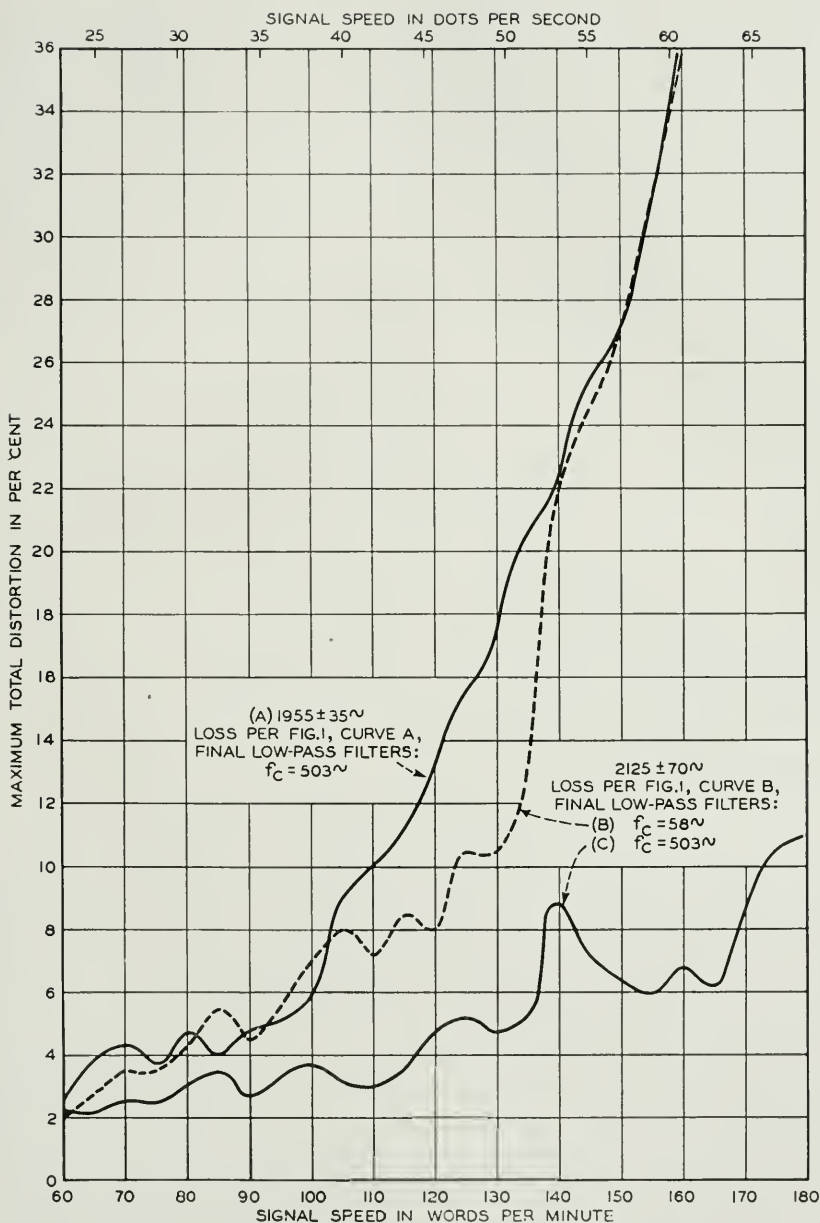
Fig. 4—Distortion vs. speed characteristics of normal- and wide-band frequency-shift arrangements.

have zero bias while transmitting the test sentence because the characteristic of the level compensator was such that reversals could not be used to adjust for zero bias of the received signals. The distortion vs. speed characteristics of the level compensated on-off and 40B1 arrangements are given by curves A and C, respectively, in Fig. 5. This figure indicates that at high speeds the frequency-shift arrangement is subject to somewhat greater distortion than the on-off method.

It may be argued that Fig. 5 is not a fair comparison between frequency-shift and on-off methods because of the difference in detector characteristics. In order to overcome this objection an experimental on-off arrangement was set up utilizing a linear detector and the same receiving relay as in the frequency-shift arrangement. The effective operating ampere-turns in the 255A receiving relay were kept the same for both frequency-shift and linear on-off methods of transmission. The line winding current varied from about 10 mils during spacing signals to 50 mils during marking signals, and the biasing winding current tending to move the armature toward spacing was about 30 mils for the on-off arrangement. The effective relay operating current in the frequency-shift arrangement was $+20$ mils in the marking condition and -20 mils in the spacing condition.

The distortion measurements are shown in Fig. 6. In order to compare frequency-shift with the linear on-off arrangement, consider curves A and B of Fig. 6. There is not much difference between them, but the frequency-shift characteristic shows slightly higher distortion over part of the speed range, as in Fig. 5.

In these tests the channel loss characteristic used was that of Fig. 1, curve A.

Two-Source and One-Source Two-Band Arrangements

The same linear detector, receiving relay, and effective relay operating current were used for these two-band arrangements as for the frequency-shift arrangement. Curves C, D, and E of Fig. 6 show the speed characteristics of various two-source and one-source two-band arrangements in which the marking and spacing paths had loss characteristics similar to curve A of Fig. 1. Curve C of Fig. 6 applies to the arrangement using two oscillators and no limiter and does not differ greatly from the characteristic for the on-off method, curve B. Curve D applies to the arrangement using two oscillators and limiter, and shows greater distortion than curve C because of modulation products arising in the limiter between the sidebands of the marking and spacing carriers. This type of interference was due to discontinuities in phase of the carrier wave at the signal transitions, and was eliminated by the use of a frequency modulated oscillator in place of the two independent oscillators, as indicated by curve E. The latter is somewhat similar to curve C on Fig. 4 measured on the frequency-shift arrangement with the wide
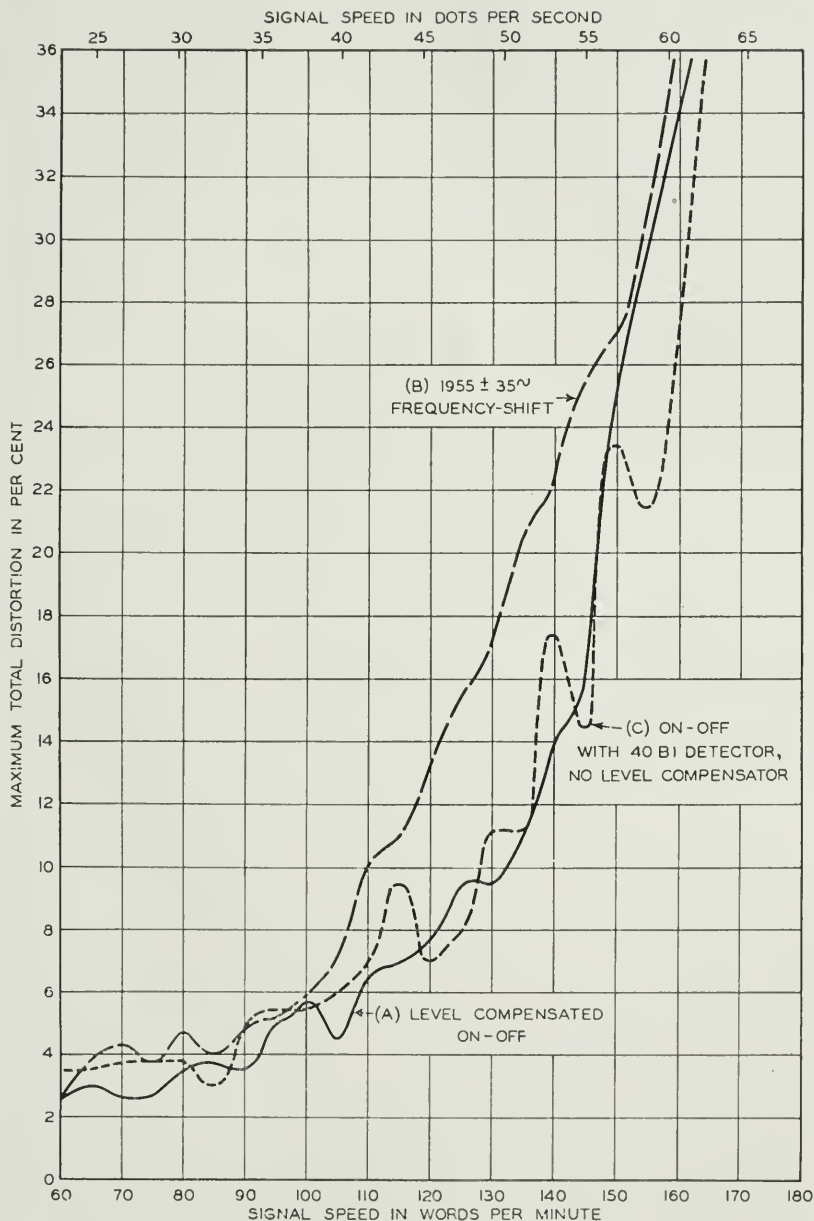
Fig. 5—Distortion vs. speed characteristics of frequency-shift and non-linear
on-off arrangements.
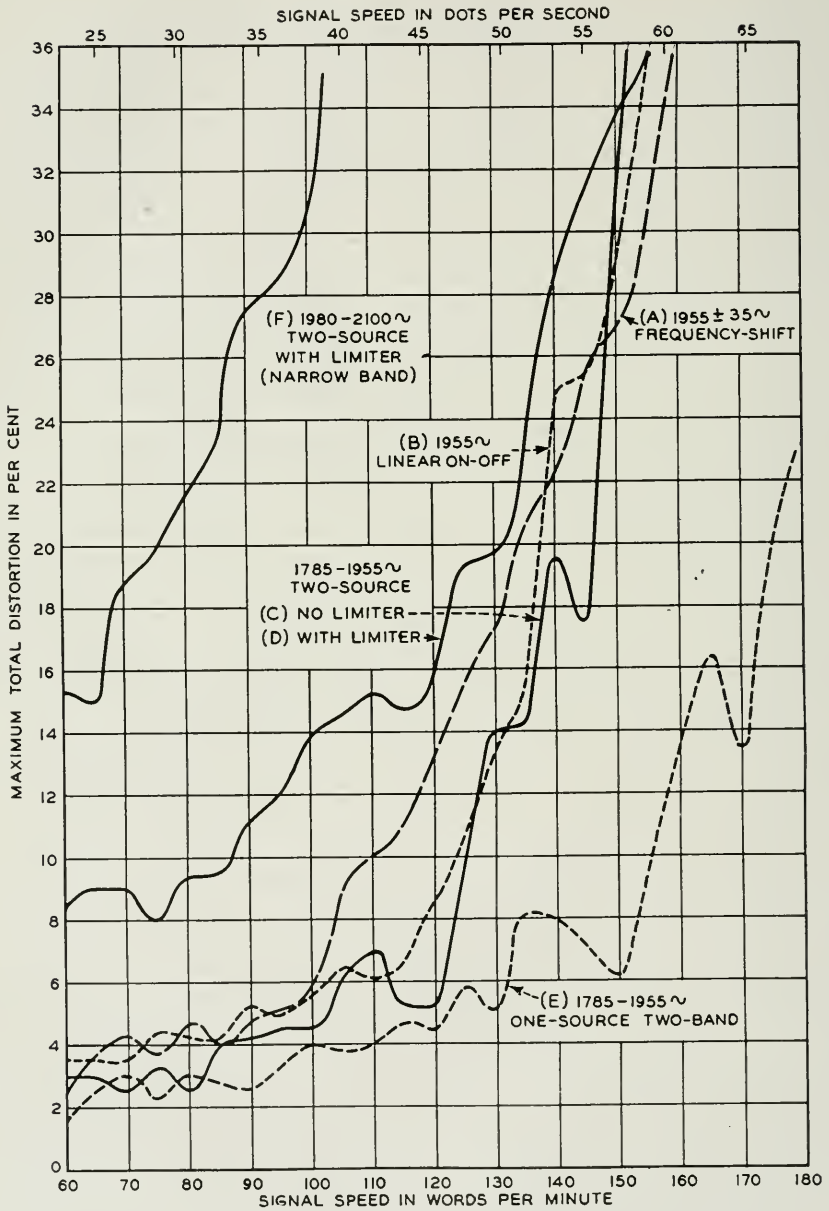
Fig. 6—Distortion vs. speed characteristics of frequency-shift, linear on-off, two-source and one-source two-band arrangements.

band, except that the distortion at the higher speeds is greater with curve E of Fig. 6 because sideband components in the middle portion of the total frequency band were considerably attenuated by the channel filters of the one-source two-band arrangement, but not by those of the wide-band frequency-shift arrangement. The distortion vs. speed characteristic of the two-source arrangement having paths with loss characteristic similar to curve C of Fig. 1, is shown by curve F of Fig. 6. The large distortion is due to the narrower sidebands transmitted and to the combined use of two independent oscillators and a limiter, the effect of which is discussed above.

Single-Sideband Arrangement

The same linear detector, receiving relay, and effective relay operating current were used for the single-sideband tests as for the linear on-off tests.

An ideal single-sideband arrangement should operate at twice the speed of the on-off arrangement for the same pass band, if the quadrature component[2] is eliminated. The cost of a phase discrimination method of reception[1] for this purpose would probably be prohibitive in practice. If the quadrature component is allowed to remain, it is a principal cause of distortion, so that the single-sideband method gives only a slight increase in speed. The effect of the quadrature component on telegraph distortion can be reduced[2] by the transmission of a certain amount of spacing carrier current. Curve B of Fig. 7, measured with a spacing current 6 db below the marking current, shows less distortion than curve A of Fig. 7, measured with no spacing current; and, in the range of speeds investigated, does not differ greatly from curve C, taken on the linear on-off arrangement without channel filters. Thus, it is apparent that the single-sideband arrangement is capable of higher speeds, for a given distortion and band width, than the other arrangements here considered.

Tests of Carrier Frequency Variations

When a carrier telegraph circuit contains a radio or carrier telephone link, some instability may occur in the average received carrier frequency. In order to investigate the effect of varying the mean carrier frequency, the carrier supply frequency was varied as a matter of convenience. Since the signals were transmitted through both sending and receiving channel filters the effects observed were doubtless about twice as bad as if only the received carrier frequency had been varied, except perhaps in the frequency-shift arrangements where the discriminator produced a large effect.

Distortion at 60 Words per Minute

In Figs. 8 and 9, the distortion obtained over the various arrangements is shown as a function of carrier frequency variation from the nominal value,

when the speed was 60 w.p.m.   The marking and spacing frequencies of the
two-source, one-source two-band, and frequency-shift arrangements were
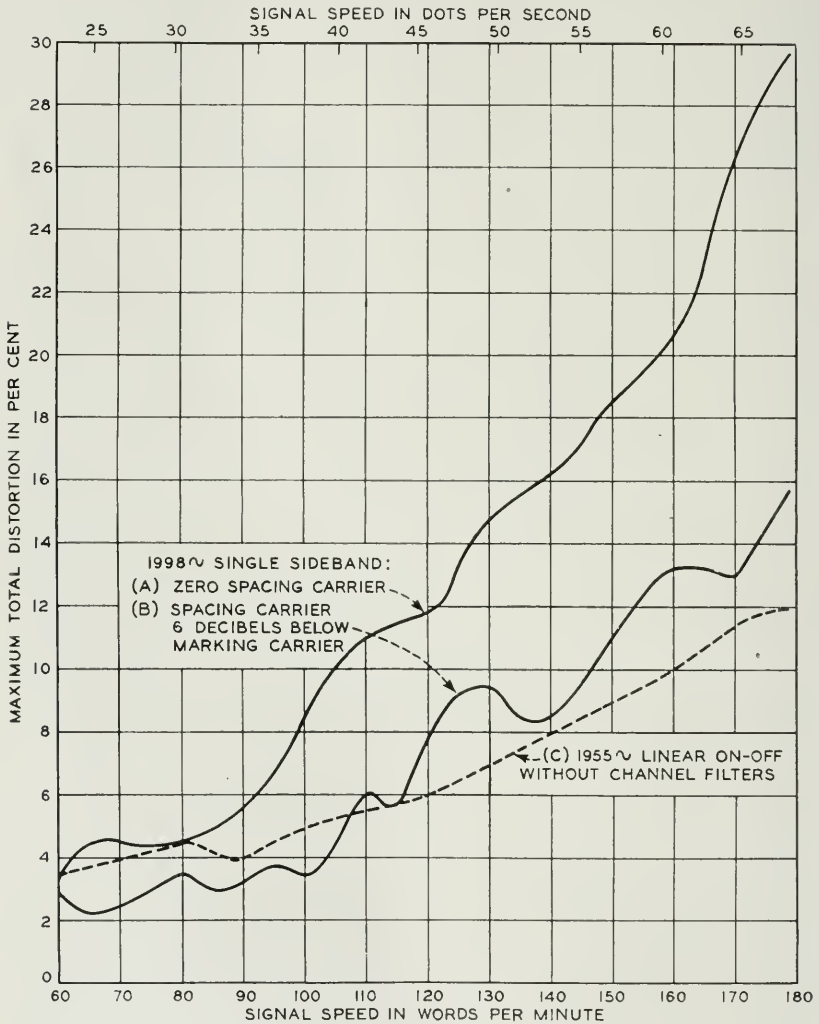
Fig. 7—Distortion vs. speed characteristics of single-sideband arrangements and
on-off arrangement without channel filters.

both varied by the same amount from their normal values without changing
their relative separation.   It is evident from Fig. 8 that the arrangements
ranked in the following order as regards the permissible carrier frequency
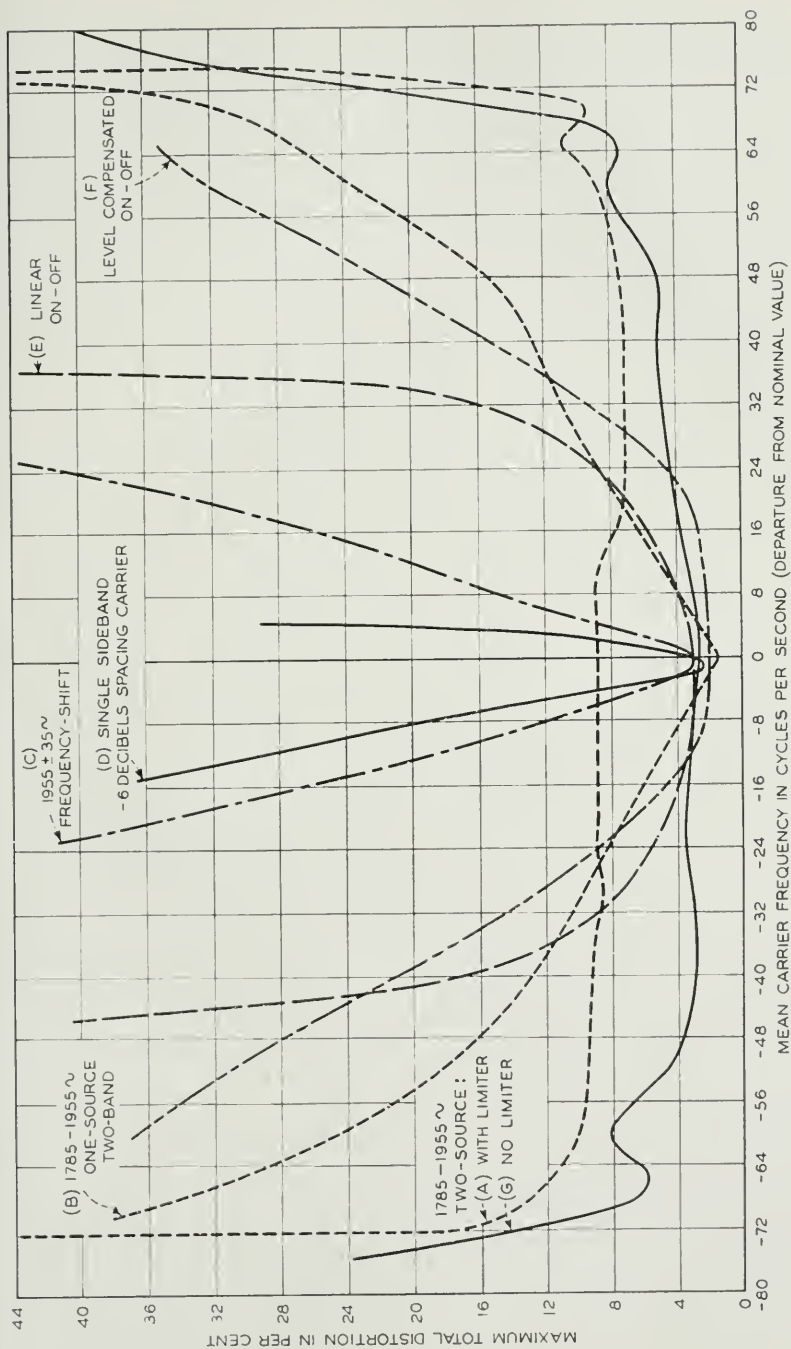
Fig. 8—Distortion vs. mean carrier frequency characteristics of normal-band arrangements at 60 w.p.m. (23 d.p.s.).

variations when the distortion was about 20 per cent: two-source without limiter on a par with two-source with limiter, one-source two-band with limiter, level compensated on-off, linear on-off, frequency-shift, and single-sideband with −6 db spacing carrier. This comparison includes only those arrangements using the same type of filter (with loss per curve A, Fig. 1). Figure 9 shows the results obtained by making similar tests on the wide-band frequency-shift (with loss per curve B, Fig. 1) and on the narrow-band two-
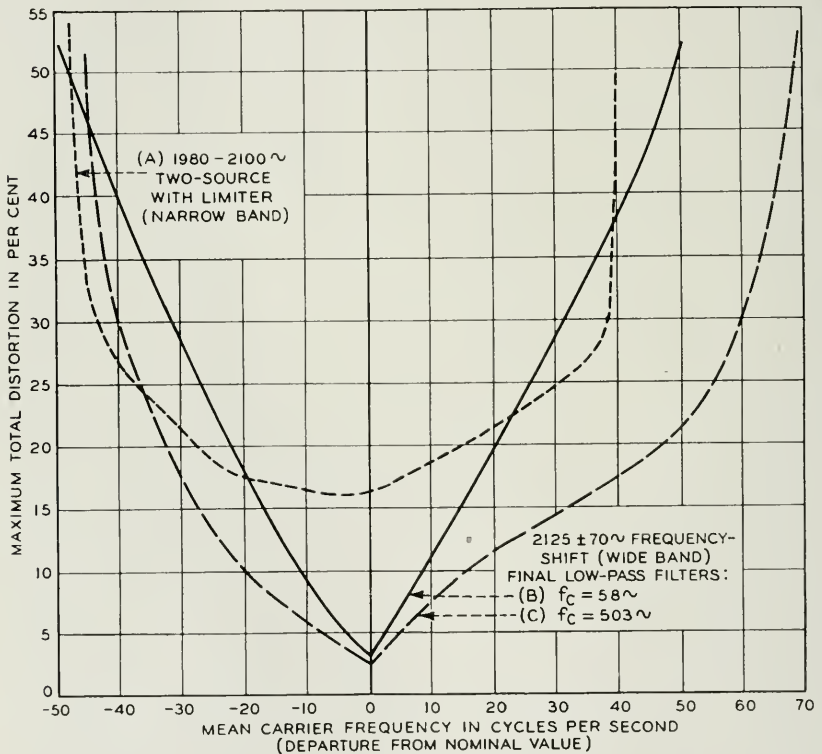
Fig. 9—Distortion vs. mean carrier frequency characteristics of narrow- and wide-band arrangements at 60 w.p.m. (23 d.p.s.).

source arrangements (with loss per curve C, Fig. 1). The distortion curves for the wide-band frequency-shift arrangement in the latter figure are more favorable than the curve for the normal-band frequency-shift arrangement shown in Fig. 8 because the wide-band arrangement had a discriminator with only half the slope of that used in the normal-band arrangement. The wide-band arrangement with the higher cut-off low-pass filter could tolerate a greater change in carrier frequency than the arrangement with the lower

cut-off low-pass filter because of the steeper wave front of the signals delivered by the former low-pass filter, resulting in a lower signal bias when the detected amplitudes of the marking and spacing signals differed.

When the average carrier frequency of a two-source arrangement was varied, the marking and spacing frequencies moved toward one side of their respective pass bands. As they approached the cut-off frequencies of the band-pass filters, telegraph distortion occurred due to suppression of the carrier and adjacent components. The narrow-band two-source arrangement was more sensitive to variation in the average carrier frequency than the normal-band two-source arrangement, and the reason is obvious.

Bias at 60 Words per Minute

When tested with reversals at 23 d.p.s. (corresponding to 60 w.p.m.), the arrangements also ranked in the same order from the bias standpoint as from the distortion standpoint, as the mean carrier frequency was varied. In Figs. 8 and 9 the distortion due to carrier frequency variation consisted mainly of bias, except in the two-source arrangements, where the received marking and spacing pulses were substantially equal on reversals so that there was little bias. (Bias measurements referred to here and below have not been shown graphically in order to save space.)

The on-off, two-source, and one-source two-band arrangements were fairly insensitive to carrier frequency variations since the loss vs. frequency characteristics of the channel filters changed but slowly near the middle of the transmission band. Since the frequency-shift arrangement had a discriminator which was sensitive to frequency changes, drifting of the average carrier frequency resulted in a rise in detected current in one half of the push-pull detector and a reduction thereof in the other half, thus causing serious bias in the differentially combined rectified waves, since no frequency compensator was provided. It is outside the scope of this paper to describe such a compensator, but it is no more complicated than the level compensator used with an on-off arrangement. In the single-sideband arrangement the carrier was located at a point on the filter loss vs. frequency characteristic where the slope was steep. Consequently small frequency changes produced large amplitude variations in the operating current of the receiving relay. Since there was no compensating change in the biasing current of the relay, large bias variations resulted from small changes in carrier frequency.

Distortion at 120 Words per Minute

Figures 10 and 11 give distortion for the arrangements at 46 d.p.s. or 120 w.p.m. when the mean carrier frequency was varied. The arrangements ranked in the following order when the distortion was 20 per cent: two-source without limiter, two-source with limiter, linear on-off, one-source two-band,
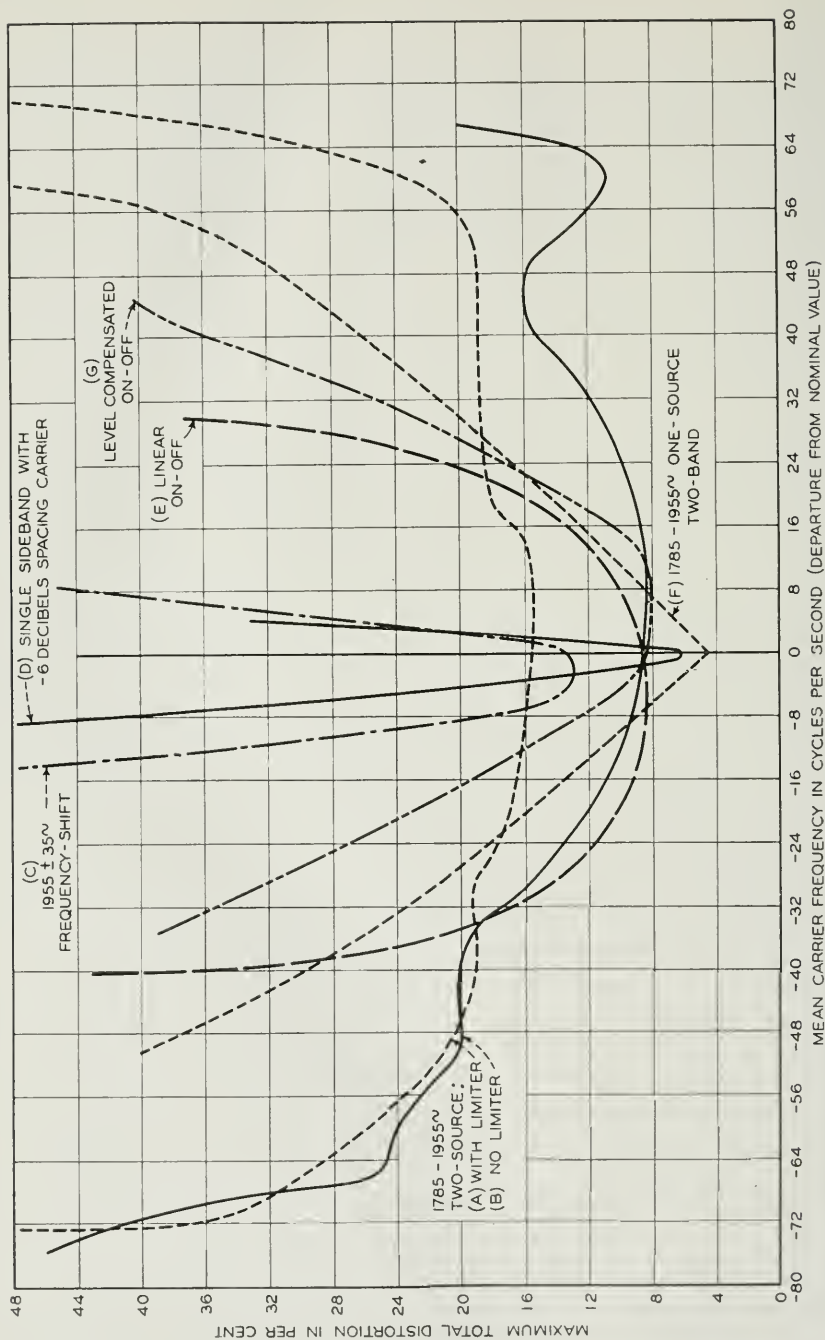
Fig. 10—Distortion vs. mean carrier frequency characteristics of normal-band arrangements at 120 w.p.m. (46 d.p.s.).

level compensated on-off, frequency-shift, single-sideband with −6 db spacing carrier.

In both Figs. 8 and 10 there appears to be considerable difference between the distortion vs. carrier frequency characteristics of the two-source and one-source two-band arrangements with limiter. The two-source arrangement was better for large carrier frequency variations, and the one-source two-band arrangement was better for small carrier frequency variations.
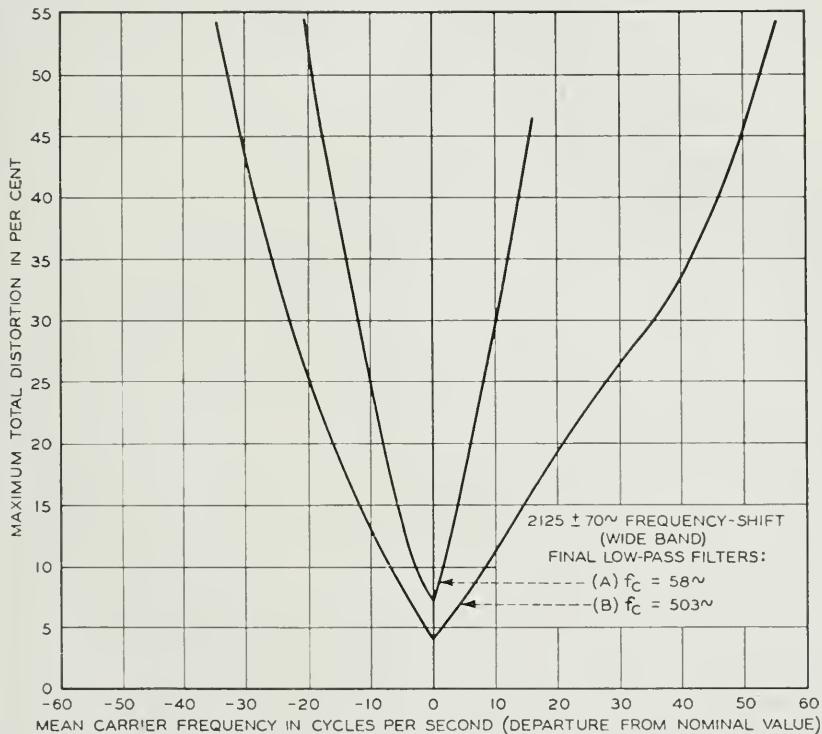
Fig. 11—Distortion vs. mean carrier frequency characteristics of wide-band frequency-shift arrangements.

The difference between the performance of these two arrangements was partly due to interference between sidebands in the two-source arrangement as previously mentioned, and partly due to the difference in the sending arrangements used. In the case of the two-source sending arrangement shown in Fig. 2, the sidebands of the marking and spacing paths were separated by the sending filters, and a small shift in carrier frequency affected each branch similarly, and relatively little bias or distortion resulted. In the case of the one-source two-band method also shown in Fig. 2, the side-

bands of the marking and spacing frequencies were not completely separated by the sending filters and a shift in carrier frequency affected the two sets of sideband components dissymmetrically, thus causing bias and distortion.

Bias at 120 Words per Minute

The bias resulting from carrier frequency change in the various arrangements at 46 d.p.s. or 120 w.p.m. was in general worse than at half this speed because the received wave shape was more rounded. The distortion due to carrier frequency variation, as shown in Figs. 10 and 11, consisted mainly of bias, except in the two-source arrangements, as explained above.

Other Considerations Relating to Single-Sideband Arrangement

The single-sideband arrangement with −6 db spacing carrier was found to have the lowest distortion at high signal speeds. Consequently it was thought desirable to study this arrangement further in order to see what might be done to improve its stability during carrier frequency variations, and to select an optimum location for the average carrier frequency. First, curve A of Fig. 12 was plotted showing the distortion at 160 w.p.m. resulting from a change in carrier frequency. Then it was assumed that a level compensator might be provided. In order to simulate the effect of such a device without actually constructing one, the line loss was adjusted manually to keep the r.m.s. marking carrier power constant at the receiving filter output. Curve B of Fig. 12 shows an improvement in the change in distortion vs. carrier frequency under this condition. Large variations in bias still persisted in spite of level compensation, due to variations in shape of the envelope of the received carrier signals caused by various amounts of quadrature component and sluggish in-phase component[1] depending on the location of the carrier frequency. In order to simulate the effect of a level compensator providing automatic bias adjustment, the relay bias current was adjusted to give zero receiving bias at each setting of the carrier frequency, and also the receiving filter output was maintained at a constant value as before. The results are given by curve C in Fig. 12 which shows a considerably increased tolerance to carrier frequency changes when the arrangement was stabilized in this manner. This curve indicates that the carrier frequency could be increased about 17 cycles before the distortion started to increase rapidly, and could be reduced about 5 cycles before a gradual increase in distortion began to appear.

It has been shown that for a certain ideal filter[1], a suitable location of the carrier frequency for the single-sideband method is at a point either at the upper or lower side of the band where the loss is 6 db with respect to that in the middle of the band. It can be demonstrated that the envelope of the received wave is the same when the carrier frequency is set at either of these

locations, if the filter characteristic is symmetrical about the midband frequency, if the pass band is narrow, and if the carrier frequency is high compared to the dot speed, as in the arrangement tested. Consequently there was no point in duplicating measurements for carrier locations at the lower edge of the band except perhaps to discover the second order effect of slight asymmetry in the channel filters. 1998 cycles is at the right-hand 6 db point of the filter characteristic given by curve A of Fig. 1; and as there is not much choice in the region from 5 cycles below to 17 cycles above this
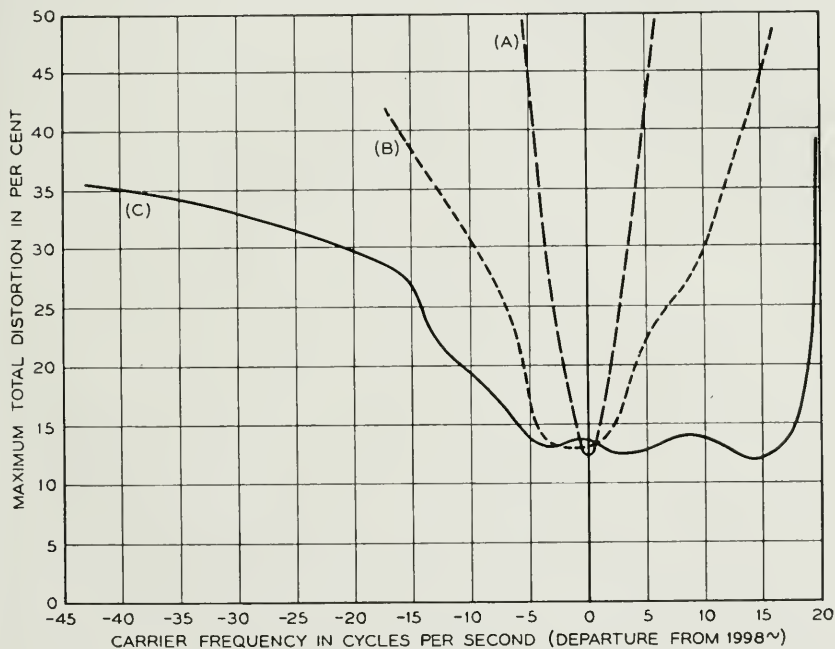
Fig. 12—Distortion vs. carrier frequency characteristics of 1998-cycle single-sideband arrangement using spacing carrier current 6 db below marking current at 160 w.p.m. (61 d.p.s.).

A: Fixed adjustments. B: Constant marking output level from receiving filter, and fixed relay bias current. C: Same as B, except receiving signal bias adjusted to zero by varying receiving relay bias current while transmitting reversals.

frequency, according to curve C of Fig. 12, it was considered satisfactory to locate the carrier at 1998 cycles for certain other single-sideband tests covered in this paper, although a slightly higher or lower value would probably have done about as well. As previously indicated, the reason for assuming a spacing carrier current 6 db below the marking value for single-sideband tests was to reduce the distortion due to the quadrature component[2]. Still less effect from the latter would have existed if the spacing carrier current had been further increased, but this would have reduced the difference be-

tween the marking and spacing current amplitudes on the line, causing a corresponding reduction in sideband power. If the quadrature component had been completely eliminated it is possible that 1998 cycles would have been found to be a more favorable carrier location than some of the other closely adjacent frequencies. However, no attempt was made to design the filters for the theoretical single-sideband requirement that the transfer admittance of the vestigial sideband should be complementary[1] to that of the other sideband near the carrier frequency.

Tests of Received Level Variations

Any transmission path is likely to have level variations caused by temperature and weather changes in case of wire lines and by fading in case of radio links. Each arrangement was therefore tested for susceptibility to level changes by varying the artificial line over a wide range. As the artificial line was made of resistances, the effect of equal fading over the entire frequency range was thereby simulated.

Distortion at 60 Words per Minute

In Fig. 13 is shown the total distortion at 60 w.p.m. vs. level change for different arrangements using the same type of filter. As may be seen from Fig. 13, the arrangements which had the same loss characteristic ranked in the following order at 20 per cent distortion and 60 w.p.m. as regards stability when the line level was varied: two-source with limiter on a par with one-source two-band, frequency-shift, level compensated on-off, two-source without limiter, linear on-off, and single sideband with −6 db spacing carrier. The range of levels over which the first three arrangements mentioned above were stable, in the absence of interference, was largely a function of the range of the limiter, which was over 80 db. The on-off arrangement including a level compensator had a range of approximately 40 db, but it should be remembered that a level compensator is effective only for level changes slow compared to the signal speed. The range of the two-source arrangement without limiter depended on the accuracy with which the marking and spacing halves of the receiving circuit were balanced. In the arrangement tested this range was about 30 db. The linear on-off arrangement without level compensator was very sensitive to level changes because the operating current in the receiving relay varied without a compensating variation in the bias current. The single-sideband arrangement had greater sensitivity than the linear on-off arrangement because a variation of 6 db in the line current of the single-sideband arrangement was accompanied by a 15 db variation in relay operating current due to the use of increased gain and a large grid bias in the receiving d-c amplifier.
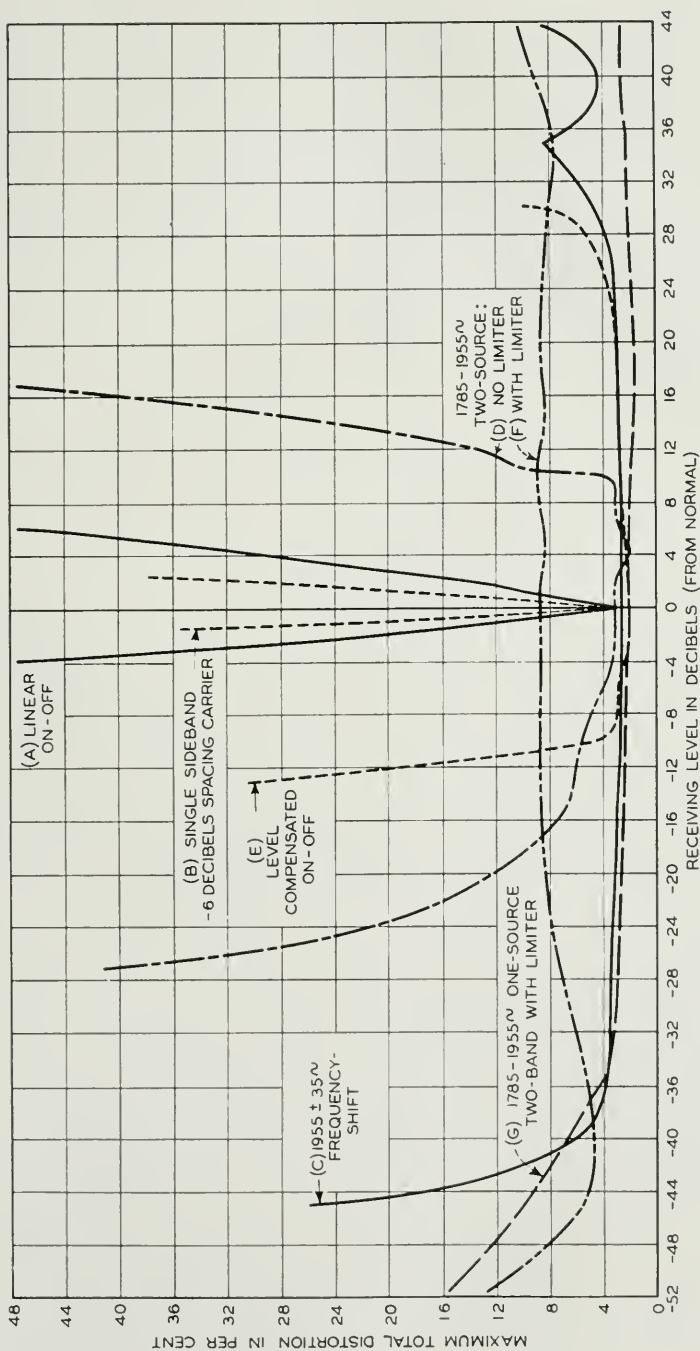
Fig. 13 — Distortion vs. receiving level characteristics of normal-band arrangements at 60 w.p.m. (23 d.p.s.).

Just as the line level range of an on-off arrangement may be extended by the use of a level compensator, so it seems reasonable to expect that a similar improvement could be obtained by using level compensators on the linear on-off and single-sideband arrangements.   However, when a level compensator is applied to such arrangements, it must be a slow acting device in order not to cause characteristic distortion.   When large and rapid level changes are frequent, as may occur on a radio circuit due to fading, an arrangement including a fast acting device like a limiter is preferred, such as a two-source, one-source two-band, or frequency-shift arrangement.

The effect of level variations on the distortion of the narrow-band two-source arrangement and of the wide-band frequency-shift arrangement with and without low cut-off low-pass filter is shown in Fig. 14.   At 60 w.p.m. the wide-band frequency-shift arrangement with low cut-off low-pass filter could
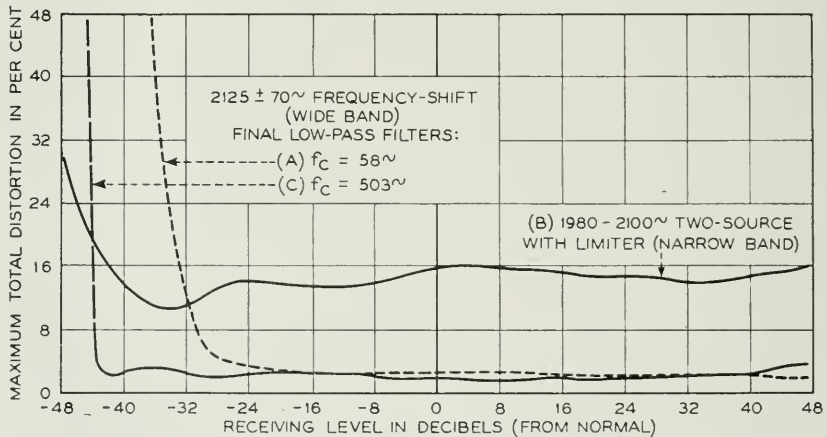
Fig. 14—Distortion vs. receiving level characteristics of narrow- and wide-band arrangements at 60 w.p.m. (23 d.p.s.).

tolerate the nearly same level change as the normal-band frequency-shift arrangement if 20 per cent distortion is the limit.   When the low-pass filter of the wide-band frequency-shift arrangement had a high cut-off, the tolerance was somewhat greater due to the steeper wave front of the detected signals which rendered them less suceptible to bias caused by slight unbalance in the detector and d-c amplifier at levels below the cut-off of the limiter.   The narrow-band two-source arrangement tolerated slightly less level change than the normal-band two-source arrangement with limiter, for the same reason.

Distortion at 120 Words per Minute

When the speed was doubled the effects of level change on distortion were found to be as shown by Figs. 15 and 16, except that tests on the narrow-band two-source arrangement were omitted because of high distortion at
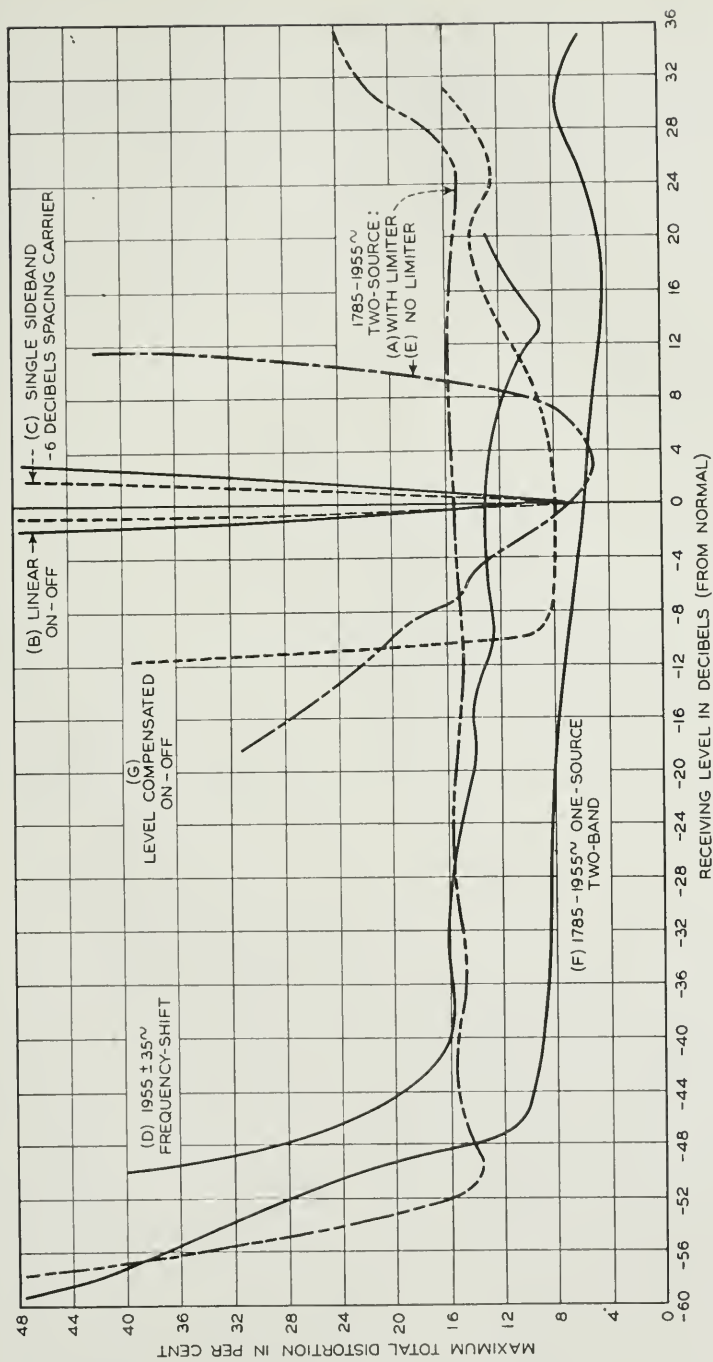
Fig. 15—Distortion vs. receiving level characteristics of normal-band arrangements at 120 w.p.m. (46 d.p.s.).

513

120 w.p.m.   The arrangements tested were found to have about the same relative susceptibility to level changes at this speed as at 60 w.p.m.

The distortion shown in Figs. 13 and 15 consisted largely of bias for the arrangements having no limiter.   The distortion shown in Figs. 13, 14, 15, and 16 for arrangements which had limiters rose faster than the absolute value of the bias as the level dropped below the cut-off of the limiter.   This was partly due to extraneous noise in the laboratory apparatus.
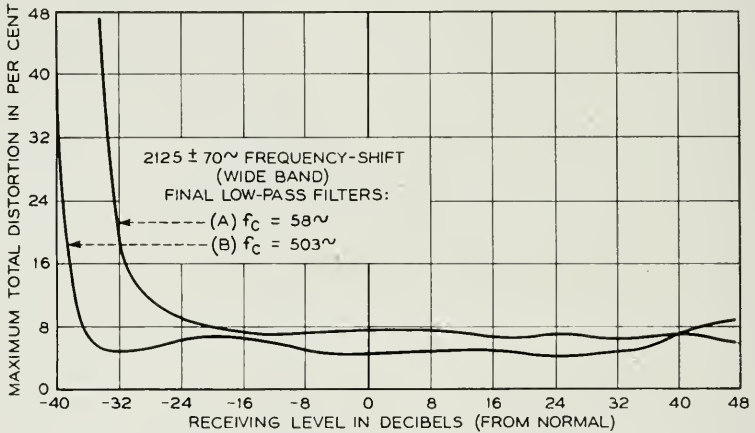
Fig. 16—Distortion vs. receiving level characteristics of wide-band frequency-shift arrangements at 120 w.p.m. (46 d.p.s.).

Selective Level Changes

When frequency-shift, two-source, or one-source two-band arrangements are operated over a radio link, selective fading may occur which affects the marking and spacing frequencies by different amounts.   Tests were made only on the two-source arrangements to measure the effect on distortion and bias of differences between received marking and spacing levels.   This effect was simulated by setting the level of one of the two carrier oscillators at different values with respect to the other, and measuring the distortion and bias for each level setting.   The distortion measurements are shown in Fig. 17 and are summarized below in Table I.

It may be seen from Fig. 17 that the distortion rose rather rapidly as the marking level was changed with respect to the spacing level and thus the distortion due to selective fading was not greatly reduced by the use of a limiter. The increase in distortion was mainly bias.

Tests of Adjacent Channel Crossfire

In order to investigate the effects on distortion due to interference from adjacent channels, certain arrangements having the same loss characteristic

were tested while upper and lower adjacent flanking channels were transmitting reversals produced by separate vibrating relays at roughly 23 d.p.s.
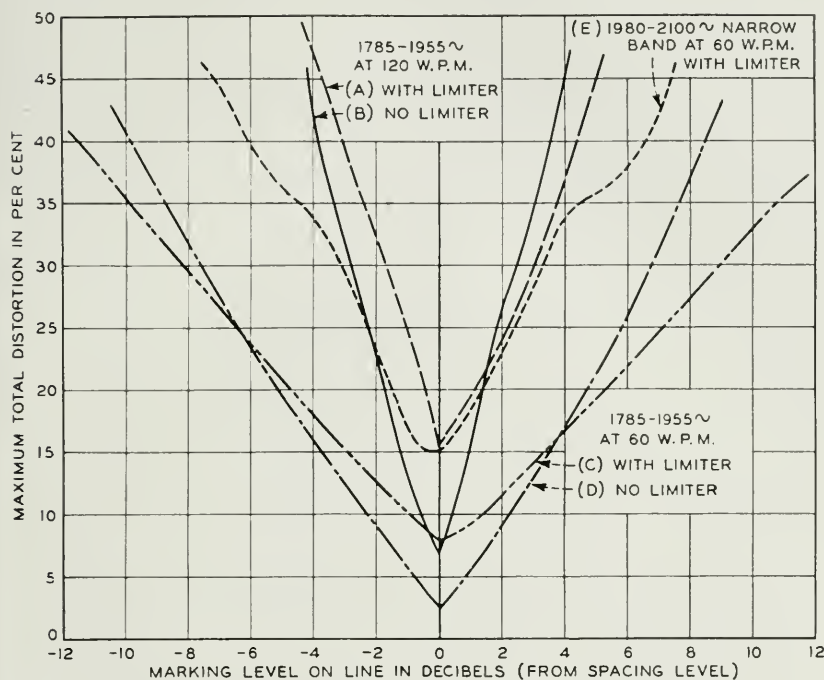
Fig. 17—Distortion vs. difference between marking and spacing level.
Comparison of two-source arrangements.

TABLE I

COMPARISON OF TWO-SOURCE ARRANGEMENTS, EFFECT OF DIFFERENCE BETWEEN MARKING AND SPACING LEVELS

Mid-band Frequencies	Lim-iter	Marking Level on Line—DB from Spacing Level			
		For 10% Distortion		For 20% Distortion	
		At 60 W.P.M.	At 120 W.P.M.	At 60 W.P.M.	At 120 W.P.M.
1785–1955~	Out	−2.2 to +2.2	−0.6 to +0.4	−5.1 to +4.7	−1.7 to +1.4
1785–1955~	In	−1.0 to +1.4	*	−4.7 to +5.3	−0.4 to +1.0
1980–2100~ (narrow band)	In	*	failed	−1.5 to +1.3	failed

* Distortion exceeded 10%.

The distortion with and without the flanking channels operating and the difference in distortion are shown in the following tables, from which it is seen that the various arrangements tested ranked approximately in the fol-

lowing order as regards their insensitivity to flanking channel crossfire: linear on-off and two-source without limiter about the same, single sideband with no spacing carrier about the same as with −6 db spacing carrier, frequency-shift.

It may be observed from Tables II and III that the interchannel crossfire obtained for the linear on-off and two-source arrangements was small, due to the location of the carrier frequencies at the centers of the channel filter bands. The crossfire obtained for the single-sideband and frequency-shift methods given in Tables IV to VI was greater, due to the location of the marking or spacing frequencies near the edges of the channel filter bands.

TABLE II

1955-CYCLE LINEAR ON-OFF ARRANGEMENT, EFFECT OF OPERATING ADJACENT 1785-CYCLE AND 2125-CYCLE ON-OFF CHANNELS OVER LINE

Signal Speed, W.P.M.	Maximum Total Distortion, Per Cent		Increase in Distortion, Per Cent
	Flanking Channels Off	Flanking Channels On	
100	6.2	6.2	0
120	7.8	8.6	0.8
140	18.0	20.0	2.0

TABLE III

1785 AND 1955-CYCLE TWO-SOURCE ARRANGEMENT WITHOUT LIMITER, EFFECT OF OPERATING ADJACENT 1615-CYCLE AND 2125-CYCLE CHANNELS OVER LINE

Signal Speed, W.P.M.	Maximum Total Distortion, Per Cent		Increase in Distortion, Per Cent
	Flanking Channels Off	Flanking Channels On	
80	3.2	4.2	1.0
100	3.5	4.5	1.0
120	7.7	8.5	0.8
140	18.5	19.7	1.2

The other arrangements previously mentioned were not tested for flanking channel crossfire. The two-source arrangement with limiter and normal band width is thought to be no worse from this standpoint than that without limiter, since a limiter usually helps in discriminating against small spurious currents. From theoretical considerations the one-source two-band arrangement with limiter is thought to be better than the two-source arrangement without limiter, using the same channel filters and carrier frequencies. The narrow-band two-source arrangement had considerably sharper cut-off filters than the normal-band arrangement, thus causing greater attenuation of frequencies outside the desired band. Consequently the narrow-band two-source arrangement is thought to be no worse than the normal-band arrangement as far as interchannel interference is concerned. The wide-

band frequency-shift arrangements are thought to be no worse than the normal-band frequency-shift arrangement from this standpoint, because of the greater separation between the sidebands of the adjacent channels. According to a study of the frequency spectra produced by abrupt or sinusoi-

TABLE IV

1998–CYCLE SINGLE-SIDEBAND ARRANGEMENT (NO SPACING CARRIER), EFFECT OF OPERATING ADJACENT 1828-CYCLE AND 2168-CYCLE SINGLE-SIDEBAND CHANNELS OVER LINE

Signal Speed, W.P.M.	Maximum Total Distortion, Per Cent		Increase in Distortion, Per Cent
	Flanking Channels Off	Flanking Channels On	
100	7.1	7.9	0.8
120	12.0	14.9	2.9
140	18.5	22.7	4.2
160	21.0	25.5	4.5

TABLE V

1998-CYCLE SINGLE-SIDEBAND ARRANGEMENT (WITH SPACING CARRIER 6 DB BELOW MARKING CARRIER), EFFECT OF OPERATING ADJACENT 1828-CYCLE AND 2168-CYCLE SINGLE-SIDEBAND CHANNELS OVER LINE

Signal Speed, W.P.M.	Maximum Total Distortion, Per Cent		Increase in Distortion, Per Cent
	Flanking Channels Off	Flanking Channels On	
100	5.7	9.2	3.5
120	8.5	12.0	3.5
140	14.0	17.7	3.7
160	14.0	18.2	4.2

TABLE VI

1955±35-CYCLE FREQUENCY-SHIFT ARRANGEMENT (WITH RELAY MODULATOR), EFFECT OF OPERATING ADJACENT 1785 ± 35-CYCLE AND 2125 ± 35-CYCLE FREQUENCY-SHIFT CHANNELS OVER LINE

Signal Speed, W.P.M.	Maximum Total Distortion, Per Cent		Increase in Distortion, Per Cent
	Flanking Channels Off	Flanking Channels On	
60	2.5	6.0	3.5
100	5.5	11.0	5.5
120	11.0	19.0	8.0
140	20.0	28.0	8.0

dal frequency-shift arrangements during transmission of reversals[12], it appears that the sinusoidal shift is better than abrupt shift from the standpoint of interference between adjacent channels, when the sending channel filter does not sufficiently attenuate undesired sideband components. But

[12] Balth. van der Pol: "Frequency Modulation", *Proc. I. R. E.*, Vol. 18, No. 7, July 1930, pp. 1194–1205.

there is probably little use in complicating the modulating arrangement to produce sinusoidal shift, merely for the purpose of simplifying the sending filter.

In the multi-channel (on-off type) voice frequency carrier telegraph used in the Bell System plant, the carrier currents of the different channels have frequencies which are odd multiples of an 85-cycle base frequency, and the channel filters have corresponding midband frequencies. Even order modulation of these carrier currents, which occurs to a certain extent in the line repeaters of the system, results in the production of interfering frequencies which are even multiples of 85 cycles. These products fall midway between the pass bands of the receiving channel filters and the loss which they encounter in these filters greatly reduces their effect. In a telegraph system having this channel frequency arrangement, but designed to operate on a frequency-shift basis, with the carrier frequencies shifted over a large portion of the channel frequency bands, even order modulation products originating in the line repeaters would, to a much greater extent, lie in frequency ranges freely passed by the receiving filters; and the effect of such interference would be correspondingly greater than in the on-off system.

Noise Tests

One way to judge the relative noise sensitivity of carrier telegraph arrangements is to subject each to measured amounts of noise on the line and then to compare the resulting signal distortions. Resistance or thermal noise was used in these tests because it is the most general kind of noise. It consists of a superposition of rapidly recurring random impulses, some of which may overlap. No tests were made using impulse noise such as caused by lightning, ignition, or sharp static, because it was thought that resistance noise tests would suffice. Impulse noise, when considered in a strict mathematical sense, consists of isolated pulses of very short duration and the component frequencies are so phased with respect to each other that their amplitudes add arithmetically at the instants of occurrence of the pulses. Atmospheric disturbances range all the way from isolated pulses to grinding static caused by dust storms which has characteristics approaching those of resistance noise. It is difficult to choose a representative type of impulse noise for testing. Another reason for not testing with impulse noise was that theoretical considerations[13] indicate there is not much difference in the advantage of frequency-shift over on-off methods whether the disturbance is of the impulse or resistance type.

In order to compare the sensitivities of the different arrangements to re-

---

[13] M. G. Crosby: "Frequency Modulation Noise Characteristics", Proc. I. R. E., Vol. XXV, No. 4, April 1937, pp. 472–514.
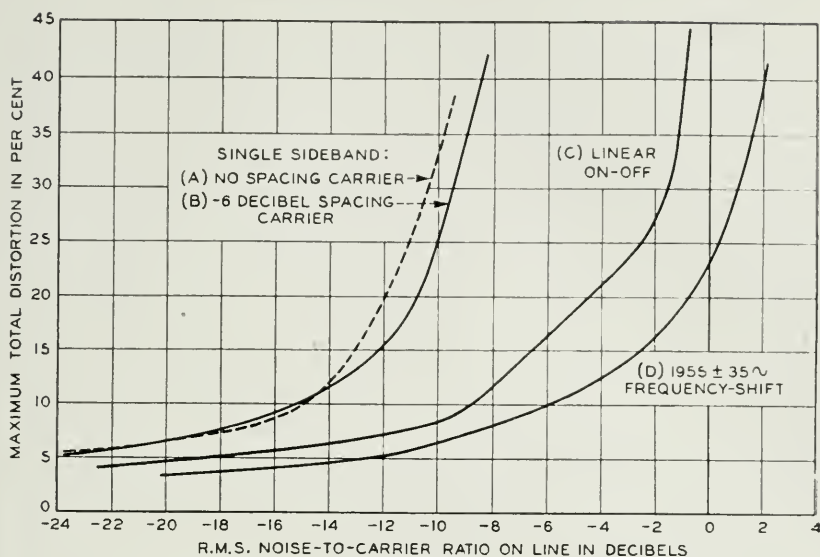
Fig. 18—Distortion vs. distortion characteristics of single-sideband, linear on-off and frequency-shift arrangements at 60 w.p.m. (23 d.p.s.).
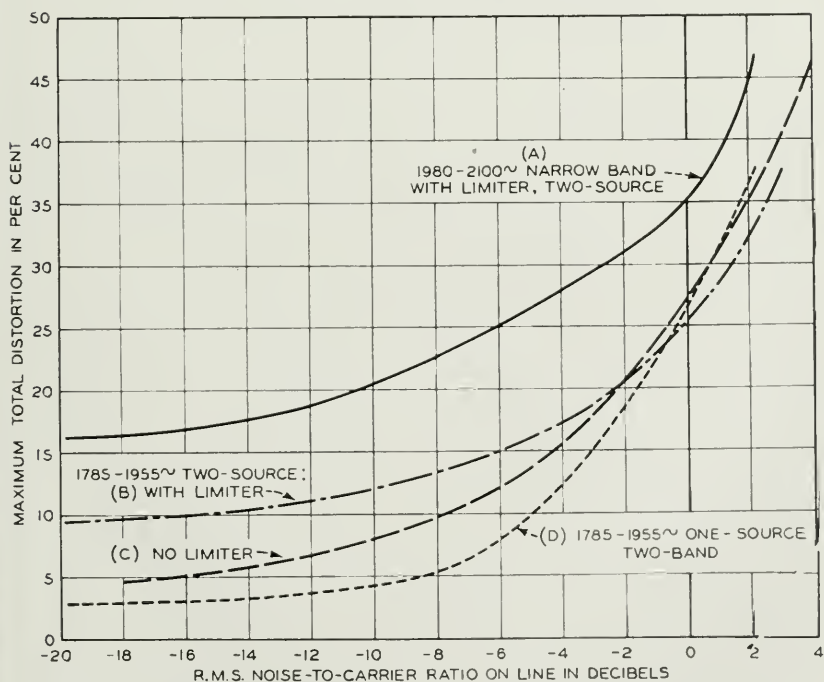
Fig. 19—Distortion vs. noise characteristics of two-band arrangements at 60 w.p.m. (23 d.p.s.).
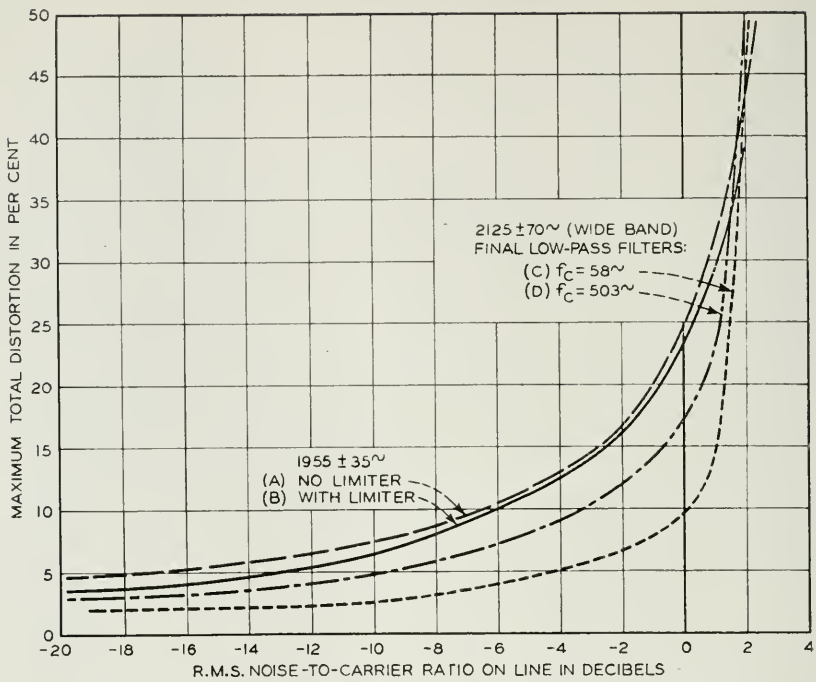
Fig. 20—Distortion vs. noise characteristics of frequency-shift arrangements at 60 w.p.m. (23 d.p.s.).
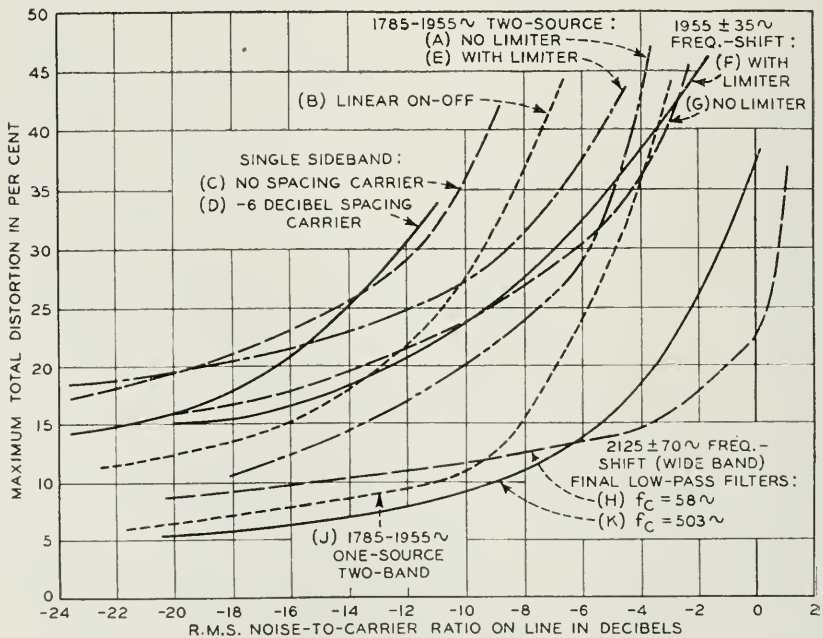
Fig. 21—Distortion vs. noise characteristics of various arrangements at 120 w.p.m. (46 d.p.s.).

sistance noise, the noise was introduced into the line through a symmetrical three-way pad, as previously mentioned. The marking carrier level at one input of the noise pad was kept at a constant value. The level of the noise current entering the other input of the noise pad was adjustable, and its r.m.s. value was measured with a thermocouple, permitting the computation of the r.m.s. noise-to-carrier ratio at the output of the three-way pad, since the loss through the pad was the same for both signal and noise. This ratio was used for abscissae in Figs. 18 to 21, inclusive, and the ordinates represent distortion for various arrangements. These curves are useful in comparing the relative noise sensitivities of the different arrangements, because the same noise source was used in all tests. The absolute noise sensitivities of the various arrangements were not known accurately because the band width of the noise source was not known exactly. The band width must have been about 3 kc. because the measured signal-to-noise power ratio of the on-off arrangement was 15 db better at the receiving filter output than on the line. Fifteen db corresponds to a power ratio of 31.6, which also should be the ratio of the band width of the noise source to the 95-cycle band width of the receiving filter. In comparing the various distortion vs. noise curves, one should note that certain arrangements have different amounts of distortion when noise is absent, which affects the comparison when noise is present.

Distortion at 60 Words per Minute

According to Figs. 18, 19, and 20 the arrangements having the same loss characteristic rank in the following order as regards their insensitivity to resistance noise at 60 w.p.m. and 20 per cent distortion: frequency-shift with limiter, closely followed by frequency-shift without limiter, one-source two-band with limiter, two-source without limiter on a par with two-source with limiter, linear on-off, single-sideband with −6 db spacing carrier, and single-sideband with no spacing carrier. All the tests recorded in Figs. 18, 19, and 20 were made on arrangements having linear receiving detectors. No noise data were taken by the writers on the level compensated on-off arrangement. Measurements made by other Bell System engineers on a similar level compensated arrangement follow the general shape of curve C of Fig. 18 for the linear on-off arrangement, except for an average displacement of about one db to the left. It is not known whether the difference in performance of the level compensated on-off arrangement was due to the difference in detector characteristics or to a difference in measuring technique.

The frequency-shift arrangements tested were all less sensitive to noise than the on-off arrangement. When a limiter was not used, this difference was due mainly to differential recombination of the rectified output currents

of the two branches of the discriminator[14]. When a limiter was used and the noise was small compared to the carrier current, the limiter theoretically should have reduced the noise-to-carrier ratio at least 3 db. Therefore, one might expect curves A and B of Fig. 20 to be separated horizontally at least 3 db in the region of low noise and to come together as the noise approaches zero. However, since curves A and B are fairly flat in the region of low noise, small errors in distortion measurement may have caused appreciable errors in the horizontal separation between the curves. When noise is not small compared to signal current, it is expected from theory that a limiter would give only a small reduction in distortion caused by noise on a frequency-shift arrangement. This appears to be verified experimentally by curves A and B of Fig. 20.

The main purpose of testing the wide-band frequency-shift arrangements was to compare their sensitivity to noise with that of the normal-band frequency-shift arrangement with limiter. According to Fig. 20, curves B and D, at 60 w.p.m. the tolerance to noise interference in the wide-band arrangement with low-pass filters having a cut-off frequency of about 503 cycles was 2.7 db greater than that in the normal-band arrangement at 10 per cent distortion and 1.3 db greater at 20 per cent distortion. This improvement was unexpected because the wider band admitted more noise. The improvement must have been due to the smaller distortion in the wide-band arrangement when no noise was present. The theoretical difference in noise tolerance between these two frequency-shift arrangements, which had a two-to-one ratio of frequency band width and of frequency swing, would have been 6 db at low noise levels if low-pass filters had been used at the detector outputs of the wide-band channel to cut off components in the detected current which are higher in frequency than those transmitted by the normal-band arrangement[13]. However, the low-pass filters used in the tests had a cut-off well above this value in order to permit signaling at very high speeds. Consequently, there was some unnecessary noise passed by these filters which accounts for an increase in noise tolerance of less than 6 db at 60 w.p.m.

In order to verify the fact that low cut-off low-pass filters improve the noise tolerance[13] of the wide-band frequency-shift arrangement, the 503-cycle cut-off filters were replaced by filters having a cut-off at about 58 cycles. According to Fig. 20, curves B and C, at 60 w.p.m. the tolerance to noise interference in the wide-band arrangements was then 6.1 db greater than in the normal-band arrangement at 10 per cent distortion, and 2 db greater at 20 per cent distortion.

[14] J. R. Carson and T. C. Fry: "Variable Frequency Electric Circuit Theory with Application to the Theory of Frequency Modulation", *Bell Sys. Tech. Jour.*, Vol. XVI, No. 4, October 1937, pp. 513–540.

The superiority of the two-source arrangements over the on-off arrangement may be accounted for by differential recombination of the detected waves, and by the greater amount of sideband power transmitted with two separate carriers.

As shown in Fig. 19, the tolerance to noise of the narrow-band two-source arrangement at 60 w.p.m. was less than that of the normal-band two-source arrangement with limiter because of the greater distortion obtained when operating at this speed without noise. Also, the two-source arrangement illustrated by curves B and C in Fig. 19 appears to have been made worse at low noise values by use of a limiter. Actually the limiter probably did reduce the noise power, but the improvement in distortion obtained thereby was greatly outweighed by the distortion increase which occurred without noise when the limiter was added, as explained above under the heading "Distortion vs. Speed Tests", and illustrated in curves C and D of Fig. 6.

The single-sideband arrangements were both more sensitive to noise than the on-off arrangement. In the case of the single-sideband arrangement with no spacing carrier the presence of quadrature[2] component increased the marking bias; and in order to have zero bias on reversals it was necessary to increase the d-c bias current in the receiving relay, which reduced the effective marking current and therefore made the arrangement more susceptible to interference occurring during the marking intervals. In the case of the single-sideband arrangement with −6 db spacing carrier the amount of sideband power transmitted was less than in the on-off arrangement and consequently noise was more troublesome.

Distortion at 120 Words per Minute

Further tests were conducted at 120 w.p.m. on the arrangements having linear receiving detectors, and results are shown in Fig. 21. The increase in speed caused an increase in distortion on all arrangements at a given noise level, but the increase in speed has less effect on the single-sideband arrangements than on the others having the same loss characteristics, at medium and high noise levels. On account of the narrower sidebands transmitted on the two-source, normal-band frequency-shift, and on-off arrangements, an increase in speed was accompanied by a greater decrease in amplitude of received signal and therefore by a greater increase in noise-to-signal ratio than in the case of the single-sideband arrangements, each of which used a wider sideband. Consequently, at a given noise-to-signal ratio, the speed increase was accompanied by a greater distortion increase on the former group of arrangements than on the latter.

As was previously found at 60 w.p.m. the frequency-shift arrangement with limiter was less sensitive to noise than the arrangement without, when

the noise was small.   But the reduction in distortion was negligible, because the distortion was small in either case.

In Fig. 21, a comparison of curve F, applying to the normal-band frequency-shift arrangement with limiter, and curve K, applying to the wide-band frequency-shift arrangement with 503-cycle low-pass filter, shows that at 20 per cent distortion the increase in band width and frequency swing resulted in a 9 db improvement in noise-to-carrier ratio.   This improvement may be explained on the same basis as the 1.3 db improvement obtained at 60 w.p.m. and was larger due to the higher distortion of the normal-band arrangement when operating at 120 w.p.m.   When the cut-off frequency of the low-pass filter of the wide-band frequency-shift arrangement was changed from 503 cycles to 58 cycles, this improvement in noise-to-signal ratio was increased to 11.6 db.   It is apparent that tolerance to noise was appreciably increased by this change of filters but, of course, this was accompanied by a reduction in maximum operating speed.

Noise-to-Signal Ratio at Receiving Relay

Another method that was used to measure noise-to-signal ratio on the on-off and single-sideband arrangements was as follows: With the carrier turned off, the receiving gain was increased until the receiving relay just operated on occasional noise peaks.   Then with the carrier on the line, and with the noise absent, the receiving gain was readjusted until the receiving relay again just operated.   The difference in receiving gain in these two tests was called the noise-to-signal ratio at the receiving relay.   When this method of measuring the noise-to-signal ratio was used it was found that the on-off and single-sideband arrangements had distortion vs. noise-to-carrier ratio characteristics similar to those given in Figs. 18 and 21.   However, in order to express these characteristics correctly when substituting the words "at receiving relay" for "on line" in the scale of abscissae, it was found experimentally that it was necessary to shift the curves for the on-off arrangement 7 db to the left and to shift those for the single-sideband arrangement 4 db to the left, relative to their present positions in these figures.   Such a change of scale was necessary because the noise-to-signal ratio was smaller at the output of the receiving filter than it was on the line.

Summary Of Results

Explanation of Tables VII and VIII

Tables VII and VIII have been prepared from the test data in order to compare the various arrangements for a given amount of distortion on each. Table VII describes certain of their properties for 10 per cent maximum total distortion.   In column A (see bottom line of the table) are listed the different arrangements.   In columns B and C are given the speeds in dots

per second and words per minute respectively at which these arrangements operated with 10 per cent distortion. In taking these data all other variables such as noise and level or frequency changes were absent. (Similarly in the other columns only the variable mentioned was allowed to change.) Column D gives the order of preference for arrangements having the same channel filter loss characteristic (but not necessarily the same total frequency band) on the basis that the highest speed is the most desirable. These data apply when no flanking channels were present. The one-source two-band arrangement and the single-sideband arrangement with -6 db spacing carrier were the fastest of those having the same channel filter loss characteristic, as might be expected from the widths of the transmitted sidebands. However, for a given band width, the single-sideband arrangement with -6 db spacing carrier was the fastest. If frequency band width is of paramount importance, consideration should be limited to arrangements with the same band width, which would change the order of preference.

Columns E, F, and G apply, respectively, in place of columns B, C, and D, when flanking channels were in operation. According to column G, the single-sideband arrangement lost its place in preferential rating due to interference from the adjacent channels; and the on-off and two-source arrangements were the best, as might be expected, because their carriers were located at midband.

In column H are listed the ranges of levels over which the received signal power could vary without causing more than 10 per cent distortion. The order of preference listed in column I indicates that arrangements with limiters were the most stable.

In column L are listed the ranges of carrier frequency variations which could be tolerated without causing the distortion to exceed 10 per cent. In column M it is seen that the two-source arrangements performed better than the others when the mean frequency changed.

In column P the arrangements are rated on the basis of the resistance noise which could be tolerated on the line compared to the linear on-off arrangement. For example, the frequency-shift arrangement in item 1.11 could tolerate 2.7 db more noise than the on-off arrangement for 10 per cent distortion at 60 w.p.m. According to column Q the one-source two-band arrangement performed the best in this respect.

Table VIII is arranged similarly to Table VII except that the maximum total distortion is 20 per cent throughout, and additional data are given in columns J, K, N, O, R, and S to cover speed at 120 w.p.m.

DISCUSSION

It is believed that the circuit arrangements tested were reasonably representative of those commonly used in carrier telegraph practice, so that

TABLE VII

COMPARISON OF DIFFERENT ARRANGEMENTS WHEN MAXIMUM DISTORTION IS 10% ON EACH.
UNLESS OTHERWISE STATED: CHANNEL SPACING 170 CYCLES & LOSS PER FIG. 1, CURVE A

Note: d.p.s. = dots per sec.; w.p.m. = words per min.; o.p. = order of preference based on performance.   (In practise, consideration should also be given to economic advantages, and in this respect the on-off arrangements rank high.)

Arrangement	Basis of Comparison												
	Speed						Flat Level Change		Mean Carrier Freq. Change		Signal-to-Noise Advantage on Line Compared with On-Off at 60 w.p.m.	
	No Flanking Channel			With Flanking Channels			Range at 60 w.p.m.		Range at 60 w.p.m.			
	d.p.s.	w.p.m.	o.p.	d.p.s.	w.p.m.	o.p.	db	o.p.	c.p.s.	o.p.	db	o.p.
1. Arrangements with Same Band Width												
1.1. Frequency-Shift, 1955 ± 35 Cycles, with Relay Modulator:												
1.11. With Current Limiter...............	42	110	8	36	94	5	>82	3	11.6	5	+2.7	2
1.12. Without Current Limiter............	41	107	9								+2.2	3
1.2. Frequency-Shift, 1955 ± 35 Cycles, with Limiter, Diode Modulator:												
1.21. Abrupt Frequency Change..........	44	115	7									
1.22. Sinusoidal Frequency Change.......	39	103	11									
1.3. Linear On-Off, 1955 Cycles..........	47	123	6	48	126	1	2.1	6	60	3	0	5
1.4. On-Off with 40B1 Detector, 1955 Cycles (No Level Compensator).............	48	127	4									
1.5. Level Compensated On-Off, 1955 Cycles......	50	132	3				>41	4	59	4		
1.6. Single-Sideband, 1998 Cycles:												
1.61. No Spacing Carrier.............	40	105	10	41	108	3	1.0	7	7	6	−6.3	6
1.62. Spacing Carrier −6 db..........	56	147	2	41	107	4					−6.7	7

2. Arrangements with Different Band Widths

A	B	C	D	E	F	G	H	I	L	M	P	Q
2.1. Two-Source, 1785 and 1955 Cycles:												
2.11. Without Current Limiter.........	47	125	5	47	124	2	29	5	138	1	+1.0	4
2.12. With Current Limiter.........	33	87	12				93	1	126	2	−7.3	8
2.2. One-Source Two-Band, 1785 and 1955 Cycles, with Current Limiter.........	59	155	1				>90	2	59	4	+3.7	1
2.3. Two-Source, 1980 and 2100 Cycles, 120-Cycle Spacing (Narrow-Band), 2 Bands per Fig. 1, Curve C, with Current Limiter.........	(Distortion always greater than 10% at speeds greater than 60 W.P.M.)											
2.4. Frequency-Shift, 2125 ± 70 Cycles, 340-Cycle Spacing (Wide-Band), with Limiter, Loss per Fig. 1, Curve B:												
2.41. With Final L.P. Filters, f_c = 503 cycles	66	173	*				>90	*	36	*	+5.4	*
2.42. With Final L.P. Filters, f_c = 58 cycles	47	123	*				>81	*	20	*	+8.8	*
Column No.: A	B	C	D	E	F	G	H	I	L	M	P	Q

* No order of preference given because the loss characteristics of these channels differ from those of the other channels.

TABLE VIII

COMPARISON OF DIFFERENT ARRANGEMENTS WHEN MAXIMUM TOTAL DISTORTION IS 20% ON EACH. UNLESS OTHERWISE STATED: CHANNEL SPACING 170 CYCLES & LOSS PER FIG. 1, CURVE A. (In practise, consideration should also be given to economic advantages, and in this respect the on-off arrangements rank high.)

Note: d.p.s. = dots per sec; w.p.m. = words per min.; o.p. = order of preference based on performance.

Arrangement (same as Table VII)	Speed						Flat Level Change				Mean Carrier Frequency Change				Signal-to-Noise Advantage on Line Compared with On-Off			
	No Flanking Channels			With Flanking Channels			At 60 w.p.m.		At 120 w.p.m.		At 60 w.p.m.		At 120 w.p.m.		At 60 w.p.m.		At 120 w.p.m.	
	d.p.s.	w.p.m.	o.p.	d.p.s.	w.p.m.	o.p.	Range db	o.p.	Range db	o.p.	Range c.p.s.	o.p.	Range c.p.s.	o.p.	db	o.p.	db	o.p.
1. Arrangements with Same Band Width																		
1.1. Freq.-Shift, 1955 ±35 Cycles, Relay Modulator:																		
1.11. With Current Limiter......	51	134	9	46	121	5	>84	2	>64	3	25	5	10.6	6	+3.5	1	+0.3	3
1.12. Without Current Limiter......	55	144	7												+3.3	2	−0.5	5
1.2. Freq.-Shift, 1955 ±35 Cycles, Diode Modulator:																		
1.21. Abrupt Frequency Change....	57	151	4															
1.22. Sinusoidal Frequency Change.	50	131	10															
1.3. Linear On-Off, 1955 Cycles..........	52	137	8	53	140	3	4.8	5	1.9	6	75	4	59	3	0	5	0	4
1.4. On-Off40B1 Detector, 1955 Cycles (No Level Compensator)	56	148	5															
1.5. Level Compensated On-Off, 1955 Cycles..........	56	147	5				>42	3	>41	4	85	3	44	5				
1.6. Single-Sideband, 1998 Cycles:																		
1.61. No Spacing Carrier..........	60	157	3	50	131	4									−7.7	7	−6.4	7
1.62. Spacing Carrier −6 db........	72*	190*	1	>68	>180	1	2.5	6	1.3	7	12.3	6	6.7	7	−6.4	6	−3.7	6

528

2. Arrangements with Different Band Widths

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
2.1. Two-Source, 1785 & 1955 Cycles:																		
2.11. Without Current Limiter	55	146	6	54	141	2	37	4	19.7	5	145	1	117	1	+2.0	4	+2.8	2
2.12. With Current Limiter	50	132	10				>100	1	83	2	145	1	102	2	+2.0	4	−6.2	8
2.2. One-Source, Two-Band, 1785 & 1955 Cycles, with Current Limiter	66	175	2				>100	1	>84	1	110	2	57	4	+2.6	3	+6.0	1
2.3. Two-Source, 1980 & 2100 Cycles, 120-Cycle Spacing (Narrow-Band), 2 Bands per Fig. 1, Curve C, with Current Limiter	29	76	**				>94	**	fails		44		fails		−6.2		fails	
2.4. Frequency Shift, 2125 ± 70 Cycles, 340-Cycle Spacing (Wide-Band), with Current Limiter, Loss per Fig. 1, Curve B:																		
2.41. Final L.P. Filt., $f_c = 503^\circ$	>68	>180	**				>90	**	>89	**	80	**	37	**	+4.9	**	+9.3	**
2.42. Final L.P. Filt., $f_c = 58^\circ$	52	138	**				>84	**	>82	**	42	**	14	**	+5.6	**	+11.9	**
Column No.:	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S

* Extrapolated.

** No order of preference given because the loss characteristics of these channels differ from those of the other channels.

529

fairly general conclusions are warranted. With the use of other types of filters, relays, etc., somewhat different results would doubtless be obtained.

Relative advantages and disadvantages of frequency-shift, on-off, and single-sideband methods of carrier telegraphy were deduced from measurements utilizing the same channel filters and covering a range of signalling speeds, noise, interference, and other variables.

A frequency-shift arrangement with amplitude limiter is substantially unaffected by non-selective level changes and relatively insensitive to noise currents. However, interfering currents from similar flanking channels are greater than when using the other methods. It has bias instability when the mean carrier frequency drifts. None of the arrangements tested had any compensation for drifting of the mean carrier frequency. However, it is known that automatic compensation for this may be obtained by various special methods not here described.

A channel using the on-off method is less sensitive to frequency drift, slightly faster for a given band width, and cheaper in terminal equipment than that using the frequency-shift method. The interfering currents from similar flanking channels using the on-off method are quite small. Weaknesses of the on-off method are greater sensitivity to noise and level changes. However, on good wire lines, when a level compensator[11] is used, these weaknesses are unimportant, and the on-off method is satisfactory. Of course, these weaknesses become important on radio circuits when noise is strong and when fades are too rapid or too severe to be overcome by the level compensator.

The greatest speed for a given band width is attainable by the single-sideband method. Unfortunately this method is poor from the standpoint of interchannel interference, is the most susceptible to noise, and, unless special compensating devices are used, is the most sensitive to carrier frequency drift and level changes.

Several arrangements were also investigated which utilized approximately double the band width of those mentioned in the preceding paragraph. Among these, the two-source method is the best of all the methods herein mentioned from the standpoint of insensitivity to changes in carrier frequency, whether or not a limiter is used. If a limiter is used this arrangement ranks well in its ability to withstand non-selective level changes. Two-source arrangements are sensitive to differences between the marking and spacing levels, but some advantage is obtained by the use of a limiter. The distortion vs. speed characteristic of the two-source method without limiter is about the same as that of the linear on-off method utilizing half the band width of the two-source method. As previously explained, the use of a limiter with this method causes distortion and materially reduces the max-

imum operating speed. From a noise standpoint the two-source arrangement without limiter has a slight advantage over the linear on-off method. When a limiter is used, the two-source arrangement is inferior to the linear on-off method due to the distortion inherent in this arrangement.

In the one-source two-band arrangement the limiter does not cause distortion, and a wider range of frequency components is transmitted than by the two-source method. For these reasons, the one-source two-band arrangement has a maximum working speed considerably higher than the two-source arrangement. The one-source two-band arrangement ranks close to the two-source method with limiter in its ability to withstand non-selective level changes, but its susceptibility to carrier frequency changes is greater.

A frequency-shift arrangement utilizing approximately the same band width as the two-source arrangement is found to have appreciably higher speed and less sensitivity to noise, but unless compensation is provided this frequency-shift arrangement is considerably more susceptible to carrier frequency changes. A further reduction in noise sensitivity may be obtained by the use of a low-pass filter at the detector output with a cut-off frequency low enough to limit the speed to that attainable with the normal-band frequency-shift arrangement.

Economic considerations should also be given due weight in selecting an optimum arrangement for a specific application. For example, the on-off method has been widely used on certain wire lines of the Bell System, because appreciable level changes are gradual and the lines are relatively quiet. For this application, the level compensated on-off method therefore gives satisfactory service with a minimum amount of apparatus, and the terminal arrangements are fairly simple and easy to maintain. On radio links subject to noise and fading, the more expensive frequency-shift and two-source methods have frequently been selected because of their greater reliability under such adverse conditions. The on-off and two-source arrangements have the advantage that common carrier generators (or oscillators) may be used for a number of channels.

Abstracts of Technical Articles by Bell System Authors

Weathering of Soft Vulcanized Rubber.[1] JAMES CRABTREE and A. R. KEMP. Two separate and distinct processes are responsible for the breakdown of soft vulcanized rubber when exposed to outdoor weathering—light-energized oxidation and attack by atmospheric ozone. The former is independent of stress and controllable to any marked extent only by incorporation of opaquing fillers. The latter affects rubber only when under stress and is checked to a considerable degree by addition of certain hydrocarbon waxes as long as the stress is static. The conditions affecting these processes have been investigated, and suggestions for accelerated aging are made on the basis of the findings.

A Note on a Simple Transmission Formula.[2] HARALD T. FRIIS. A simple transmission formula for a radio circuit is derived. The utility of the formula is emphasized and its limitations are discussed.

Applications of Thin Permalloy Tape in Wide-Band Telephone and Pulse Transformers.[3] A. G. GANZ. The properties and uses of thin permalloy tapes ranging from two mils to as little as 1/8 mil thick in tape cores are described. Typical applications covered are in transformers and non-linear coils for radar and for telephone systems. Data are given on the steady a-c. properties of thin tapes up to one megacycle. Pulse magnetization of the tape is analyzed. The available flux density range with uni-directional pulses and the effects of appropriate air gaps and of reverse magnetization between pulses are illustrated. Equations are given for flux distribution, effective permeability and loss, assuming linear magnetic properties, and convenient graphs for these characteristics are included. Simple expressions are developed for effective permeability and loss, which are approximations for the high d-c. permeability and rapid transition to saturation which characterize the permalloys.

Derivation of the Lorentz Transformations.[4] HERBERT E. IVES. The Lorentz transformations were obtained by Lorentz as a succession of *ad hoc* inventions, to reconcile Maxwell's theory with the results of experiments on moving bodies. By Einstein they were derived after a discussion of the

[1] *Indus. & Engg. Chemistry,* March 1946.
[2] *Proc. I.R.E.,* May 1946.
[3] *Elec. Engg., Trans. Sec.,* April 1946.
[4] *Phil. Mag.,* June 1945.

nature of simultaneity, and the adoption of a *definition* of simultaneity which violates the intuitive and common-sense meaning of that term. It is the purpose of this paper to show that these transformations can be derived by imposing the laws of conservation of energy and of momentum on radiation processes as developed by Maxwell's methods.

The Effect of High Humidity and Fungi on the Insulation Resistance of Plastics.[5] John Leutritz, Jr. *and* David B. Herrmann. The decrease in insulation resistance of methyl methacrylate, glass bonded mica, glass mat laminate phenolic, phenol fabric, phenol fiber, and wood flour filled phenol plastic is determined during prolonged exposure of the plastics to fungi and 97 per cent relative humidity at 25 C. The same plastics with fungi present also are exposed to 87, 76, and 52 per cent relative humidity to study their recovery, and then re-exposed to 97 per cent relative humidity. Samples with cleaned surfaces and with varnished surfaces are dried and then exposed to fungi and high humidity. The insulation resistance of a fungus network on methyl methacrylate is determined at 87, 76, and 52 per cent relative humidity.

Fungus growth occurs on all the test specimens except those with cleaned or varnished surfaces. The decrease in insulation resistance is retarded by the varnish. The degradation is due entirely to moisture. The rate of recovery is dependent on the composition and structure of the materials. None of the plastics is permanently affected by exposure to fungi and high humidity. Cleaning of surfaces and removal of moisture restore the insulation resistance to its original high value in every case. Water adsorption and absorption, not fungi, are the critical factors in the deterioration of the insulation resistance of these plastics.

The Elastic, Piezoelectric, and Dielectric Constants of Potassium Dihydrogen Phosphate and Ammonium Dihydrogen Phosphate.[6] W. P. Mason. Measurements have been made of all the elastic, piezoelectric, and dielectric constants of KDP and ADP crystals through temperature ranges down to the Curie temperatures. The piezoelectric properties agree well with Mueller's phenomenological theory of piezoelectricity provided the fundamental piezoelectric constant is taken as the ratio of the piezoelectric stress to that part of the polarization due to the hydrogen bonds. It is found that the dielectric properties of KDP agree well with the theory presented by Slater based on the interaction of the hydrogen bonds with the PO_4 ions. ADP undergoes a transition at $-125°C$ which results in fracturing the crystal. This transition cannot be connected with the H_2PO_4 hydrogen bond system

[5] *A.S.T.M. Bulletin*, January 1946.
[6] *Phys. Rev.*, March 1 and 15, 1946.

which controls the dielectric and piezoelectric properties, for these lie on smooth curves that do not change slope as the transition temperature is approached. It is suggested that two separate and independent hydrogen bond systems are involved in ADP. The transition temperature and specific heat anomaly appear to be connected with hydrogen bonds between the nitrogens and the oxygens of the PO_4 ions, while the dielectric and piezoelectric properties are controlled by the H_2PO_4 hydrogen bonds.

Nonlinearity in Frequency-Modulation Radio Systems due to Multipath Propagation.[7] S. T. Meyers. A theoretical study is made to determine the effects of multipath propagation on over-all transmission characteristics in frequency-modulation radio circuits. The analysis covers a simplified case where the transmitted carrier is frequency-modulated by a single modulating frequency and is propagated over two paths having relative delay and amplitude differences. Equations are derived for the receiver output in terms of the transmitter input for fundamental and harmonics of the modulating frequency. Curves are plotted and discussed for various values of relative carrier- and signal-frequency phase shift and relative amplitude difference of the received waves.

The results show that a special kind of amplitude nonlinearity is produced in the input-output characteristics of an over-all frequency-modulation radio system. Under certain conditions, sudden changes in output-signal amplitude accompany the passage of the input-signal amplitude through certain critical values. Transmission irregularities of this type are proposed as a possible explanation of so-called "volume bursts" sometimes encountered in frequency-modulation radio circuits. In general, it appears that amplitude and frequency distortion are most severe where the relative delay between paths is large and the amplitude difference is small.

Propagation of 6-Millimeter Waves.[8] G. E. Mueller. One step in the exploration of a new band of frequencies for communications purposes is a study of the transmission properties of the medium involved. This paper describes the methods and results of measurements of attenuation due to rainfall and atmospheric gases at a wavelength of 0.62 centimeter.

The one-way attenuation due to moderate rains at 0.62 centimeter is roughly 0.6 decibel-per-mile per millimeter-per-hour. The gas attenuation is probably less than 0.2 decibel per mile.

Vicalloy—A Workable Alloy for Permanent Magnets.[9] E. A. Nesbitt. Alloys in the region of 30 to 52 per cent iron, 36 to 62 per cent cobalt, and 4

[7] *Proc. I.R.E.*, May 1946.
[8] *Proc. I.R.E.*, April 1946.
[9] *Metals Technology*, February 1946.

to 16 per cent vanadium were investigated with the result that permanent magnet materials of unusual mechanical as well as magnetic properties were discovered. The alloys differ from most age-hardening alloys in that the gamma phase, stable at high temperatures, is dispersed in the alpha phase, stable at low temperatures, instead of vice versa.

Distribution of Sample Arrangements for Runs Up and Down.[10] P. S. OLMSTEAD. Using the notation of Levene and Wolfowitz, a new recursion formula is used to give the exact distribution of arrangements of n numbers, no two alike, with runs up or down of length p or more. These are tabled for n and p through $n = 14$. An exact solution is given for $p > n/2$. The average and variance determined by Levene and Wolfowitz are presented in a simplified form. The fraction of arrangements of n numbers with runs of length p or more are presented for the exact distributions, for the limiting Poisson Exponential, and for an extrapolation from the exact distributions. Agreement among the tables is discussed.

Radar Systems Considerations.[11] D. A. QUARLES. In the broad field of radio technology, radar (object location) systems have come to occupy a relatively new but highly specialized area. Because radar is a seeing and measuring art, it has put a special premium on short wavelength and has thus tended to accelerate greatly the already rapid trend toward higher frequencies. Moreover, many radar systems are associated with computer and servo mechanisms for automatic control purpose such as gunfire, bomb release and the like. To meet these new needs, a dozen or more highly developed fields of specialization covering such components as antennas, pulse transmitters and display devices have been created. Planning a new radar system calls for an appraisal of these component arts and for selection and balancing of component characteristics to produce an integrated system. The present paper deals with such technical considerations involved in planning an overall radar system as a background for other more detailed technical expositions of the component arts.

The Effect of Rain upon the Propagation of Waves in the 1- and 3-Centimeter Regions.[12] SLOAN D. ROBERTSON and ARCHIE P. KING. This paper presents some experimental results which show the effect of rain upon the transmission of electromagnetic waves in the region between 1 and 4 centimeters.

At a wavelength of 1.09 centimeters, the waves are appreciably attenuated,

[10] *Annals of Mathematical Statistics,* March 1946.
[11] *Elec. Engg., Trans. Sec.,* April 1946.
[12] *Proc. I.R.E.,* April 1946.

even by a moderate rain. Attenuations in excess of 25 decibels per mile have been observed in rain of cloudburst proportions.

The attenuation of waves somewhat longer than 3 centimeters is slight for moderate and light rainfall. During a cloudburst, however, the attenuation may approach a value of 5 decibels per mile.

The Advancing Statistical Front.[13] W. A. SHEWHART. From the viewpoint of general education, statistics is not simply a tool as is so often stated, but a scientific way of looking at the universe; statistical method is not something apart from scientific method but *is* scientific method in which the three steps, hypothesis, experiment, and test of hypothesis, are adjusted to allow for the fact that scientific inference is only probable. Applications of statistics in this sense are rapidly extending to all fields of pure, background, and applied reserch.

A New Crystal Channel Filter for Broad Band Carrier Systems.[14] E. S. WILLIS. A new crystal channel filter for use in broad-band carrier telephone systems is described. It requires less than two-thirds as much mounting space as the earlier design and savings in materials and manufacturing effort are realized. The savings were made possible by assembling the four crystal units in one lattice-type filter section rather than two, resulting in a reduction in the number of component coils and capacitors.

[13] *Jour. Amer. Statis. Assoc.*, March 1946.
[14] *Elec. Engg., Trans. Sec.*, March 1946.

Contributors to This Issue

J. G. FERGUSON, B.S. in Electrical Engineering, Queen's University, Canada. Northern Electric Company, 1923–1926; Bell Telephone Laboratories 1926–. During the war years, Mr. Ferguson developed various telephone, radio and radar equipments. Prior to the war he developed switching equipment for central offices and PBX's. He is currently interested in the design of No. 5 crossbar equipment.

H. J. FISHER, E.E., Cornell University, 1920. U. S. Signal Corps, 1917–1919. Western Electric Company, Engineering Department, 1920–1925. Bell Telephone Laboratories, 1925–. Mr. Fisher has been engaged in the development of toll transmission systems. Since 1940 and in his present capacity as Test Engineer his work has had to do principally with system testing equipment, and during the war with the development of radar testing equipment for the Armed Forces.

G. T. FORD, B.S. in Physics, Michigan State College, 1929; M.A. in Physics, Columbia University, 1936. Bell Telephone Laboratories, 1929–. Mr. Ford was concerned with some of the early work on thermistors, and, since 1934, has been engaged in development work on small high vacuum electron tubes.

CALVIN S. FULLER, B.S., Chicago, 1926; Ph.D., 1929. Bell Telephone Laboratories, 1930–1942. Office of the Rubber Director, 1942–44. Bell Telephone Laboratories, 1944–. Dr. Fuller has been engaged in the development of organic insulations and the application of plastics to electrical apparatus.

E. I. GREEN, A.B., Westminster College, 1915; University of Chicago, 1915–1916; Professor of Greek, Westminster College, 1916–1917; U. S. Army, 1917–1919 (Captain, Infantry); B.S. in Electrical Engineering, Harvard University, 1921; Dept. of Development and Research, American Telephone and Telegraph Company, 1921–1934; Bell Telephone Laboratories, 1934–. Mr. Green's responsibilities have had to do principally with multiplex transmission systems, and during the war with radar and other projects for the Armed Forces. In his present capacity as Assistant Director of Transmission Development he is in charge of development work on carrier telephone systems and on test equipment for toll transmission systems.

THEODORE A. JONES, University of Oregon, 1919–20; B.S., Oregon State College, 1924; M.A., Columbia University, 1928. Engineering Department, Western Electric Company, 1924–25; Bell Telephone Laboratories, 1925–. Since joining the technical staff Mr. Jones has been engaged in carrier telegraph development work.

J. P. KINZER, M.E., Stevens Institute of Technology, 1925. B.C.E., Brooklyn Polytechnic Institute, 1933. Bell Telephone Laboratories, 1925–. Mr. Kinzer's work has been in the development of carrier telephone repeaters; during the war his attention was directed to investigation of the mathematical problems involved in cavity resonators.

K. W. PFLEGER, A.B., Cornell University, 1921; E.E., 1923. American Telephone and Telegraph Company, Department of Development and Research, 1923–1934; Bell Telephone Laboratories, 1934–. Mr. Pfleger has been engaged in transmission development work, chiefly on problems pertaining to delay equalization, delay measuring, temperature effects in loaded-cable circuits, and telegraph theory.

C. W. SCHRAMM, B.S. in Electrical Engineering, Armour Institute (now Illinois Institute) of Technology, 1927. Illinois Bell Telephone Company, 1927–29. Bell Telephone Laboratories, 1929–. Mr. Schramm has been concerned with the development of carrier telephone systems for both message and program use. During the war his attention was directed to the design of radar test equipment.

I. G. WILSON, B.S. in Electrical Engineering and M.E., University of Kentucky, 1921. Western Electric Company, Engineering Department, 1921–25. Bell Telephone Laboratories, 1925–. Mr. Wilson has been engaged in the development of amplifiers for broad-band systems. During the war he was project engineer in charge of the design of resonant cavities for radar testing.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

A Study of the Delays Encountered by Toll Operators in Obtaining an Idle Trunk

By S. C. RAPPLEYE

THE aim of the Bell System is to give the fastest possible toll service consistent with costs. The aim of the Intertoll Trunk Engineer is to provide the proper number of trunks in each group to obtain that objective. His problem is to gauge the effect of his work on the overall speed of service.

Overall speed of toll service is the elapsed interval from the filing of a call until conversation starts or until there is a definite report about the called party. This overall speed includes the operating time or interval required for the operators to establish the connection; the subscriber time or interval required for the calling party to give the details of the call, for the called party to answer his telephone, etc.; and the circuit delay time or interval of waiting for a trunk to become idle. This last factor may be termed the *trunk speed* interval.

The proportion of this trunk speed to the overall speed is an important factor in determining the number of trunks to be provided. If it is a large proportion of the total, a marked improvement may be expected as a result of providing more trunks. Conversely, if the trunk speed is a small proportion of the total, the improvement to be expected as a result of providing more trunks will be small also, with a diminishing rate of improvement until the trunk speed ceases to be a factor. The trunk speed in turn depends upon three factors:

Group size—number of trunks to the called city or in the direction of the called city.

The per cent. usage—the degree to which the trunks are kept busy in carrying the load offered.

Holding time—the length of time that a trunk is in use each time it is used.

Since the trunk speed depends in part on the per cent. usage, it follows that this interval will be longer in the busiest hour when the usage is greatest and will be shorter in the hours which are less busy. Consequently, the trunk speed interval over the total day will be much less than in the busiest hour.

Purpose of Study

Earlier information, based on data assembled in Cleveland in 1929 and 1930, was formulated as a series of relationships between varying degrees of loading (in terms of busy hour per cent. use) on trunk groups of different sizes and the *overall* speed of service. These relationships were set forth in a table which was to be used as a guide to the trunk provision needed to accomplish a desired overall service result.

The table also furnished the *percent* calls encountering an NC (no circuit) condition but made no specific reference to the average *duration* of NC although from the data shown it could be inferred and demonstrated that other factors, such as operating method, operating and party delays, normally have a more pronounced influence on the total day overall speed of service than the busy hour trunk provision. That being so, as changing conditions since 1930 have affected these other factors, either in the direction of faster or slower service, the relationships in terms of overall speed of service shown in that table have become less valuable as engineering guides.

The purpose of the current study, therefore, was to improve the engineering and management tools used in determining the number and arrangement of trunks required to attain faster toll service so that the investment in facilities may be used as effectively as possible.

Study Procedure

The study was based on the premise that if the size of group, per cent. usage and holding time are known, the trunk speed can be determined and will remain constant under that particular set of conditions. With this constant known, it would then become possible to construct from analyses of overall speed of service data for groups, offices, areas or networks the going relationship between the trunk speed of service and the overall speed of service and to predict with reasonable assurance the effect on the overall speed which would be brought about by changes in the group sizes or traffic characteristics. The effect of foreseen changes in operating method, force conditions or the character of the toll traffic on the overall speed can be estimated separately and taken into consideration in determining the basis of trunk provision. With such information available, trunks can be provided where they will be most effective. This is especially important during periods of major change such as the transition from war to peacetime conditions or from the ringdown to the dial method of toll operation.

The problem was therefore to determine the average delay in securing a trunk with various sizes of groups at various levels of usage with a view to:

Stating that portion of the overall speed of service which results from inability to secure a circuit, and

Constructing engineering tables based on a preselected constant circuit delay or trunk speed of service.

Arrangements were made with several Associated Companies to furnish data for this purpose which would show:

1. The average overall speed of service on different sizes of groups under various conditions of loading.
2. The minimum average overall speed interval on these same groups at times when circuit provision was not a factor, i.e., when NC conditions were not encountered.

The speeds obtained in Item 2 were subtracted from those obtained in Item 1, the difference representing that portion of the overall speed which can be attributed to circuit delay, or the trunk speed.

In order to determine these trunk speeds it was necessary to obtain from several sources as much data as possible of the following nature:

Per cent. circuit usage, by hours, as derived from group busy timing registers on selected groups of various sizes. Hours during which the traffic over a group was handled subject to posted delay were disregarded.

The number of originating terminal calls handled over the groups during the hours corresponding to the usage data and the average speed of service on these calls. The call and speed of service data were summarized first to include all calls and then separately for calls not encountering NC. Correction was made for transfer of tickets to point-to-point positions by subtracting from the speed shown on each such ticket an interval representing the average length of time required to send a ticket to point-to-point positions in the office in which the data were obtained, provided the transfer time was included in the overall speed interval. This interval of transfer time is not properly chargeable as part of the trunk speed.

These data were obtained for trunk groups of various sizes ranging from one up to eighteen trunks. To secure a comparable amount of data for the smaller groups which handle fewer calls, it was necessary to include more of the smaller groups or to continue the record for a longer period of time on such groups.

The data for all hours of the day or evening were useful because as the volume of traffic recedes from the busy hour the data are typical of the busy hour condition of other groups engineered on a more liberal basis. The very light hours also show the minimum speed interval which can be obtained when lack of an available circuit is not a factor.

Five Associated Companies obtained data at eleven toll offices on 112 intertoll groups having 561 trunks. Approximately 17,000 calls (occurring during hours when the groups were at least 40% busy) were included.

The data were summarized by size of group and by circuit usage. Separate counts were maintained for groups with and without alternate routes and for person and station traffic. The following tabulation shows the type of data available for each point, i.e., each size of group at each level of usage.

ONE TRUNK—WITH ALTERNATE ROUTE—61–70% USE

% Use	Type	All Calls			Calls Not Encountering NC		
		Minutes	Calls	Speed	Minutes	Calls	Speed
61–65	Station	111	23		39	17	
	Person	109	38		68	32	
66–70	Station	103	25		33	21	
	Person	117	28		54	25	
Total		440	114	3.86	194	95	2.04

Average Speed—All Calls. . . . . 3.86 Mins.
Average Speed—NC not Encountered. . . . . 2.04 Mins.

Average Delay Due to NC (Trunk Speed). . . . . 1.82 Mins.

The results were plotted for each level of usage by steps of 10% as shown in Fig. 1, using a 3-point moving average to smooth out the deviations and to establish a more definite trend. Each of these curves was then redrawn in relation to the others and combined results are shown on Fig. 2.

The delay intervals indicated in Fig. 2 represent the total delay which resulted from the fact that there was no circuit available when the operator was first ready to make use of one. It includes not only the time spent in waiting for a circuit to become idle but also the time required for the operator herself to return to that call if she had engaged in some other work in the meantime. If the operator is not free to utilize the circuit as soon as it becomes available some other operator may use it for another, later call. The subsequent call is then delayed less than the average, or not at all, but the original call is delayed longer than the average. While the delays experienced by individual calls may vary considerably from the average, the data have been treated in terms of averages for engineering purposes.

MATHEMATICAL FORMULAE

The summarized data were referred to the different mathematical expressions frequently applied to trunking problems, such as the Poisson, Erlang "B" and Erlang "C" formulae. It was found that the observed average NC delays were considerably shorter than the theoretical average delays in those formulae which make allowance for variable holding times, such as would be encountered in local trunking where the average trunk use is short and the deviations from average on a percentage basis are apt to be appreciable.
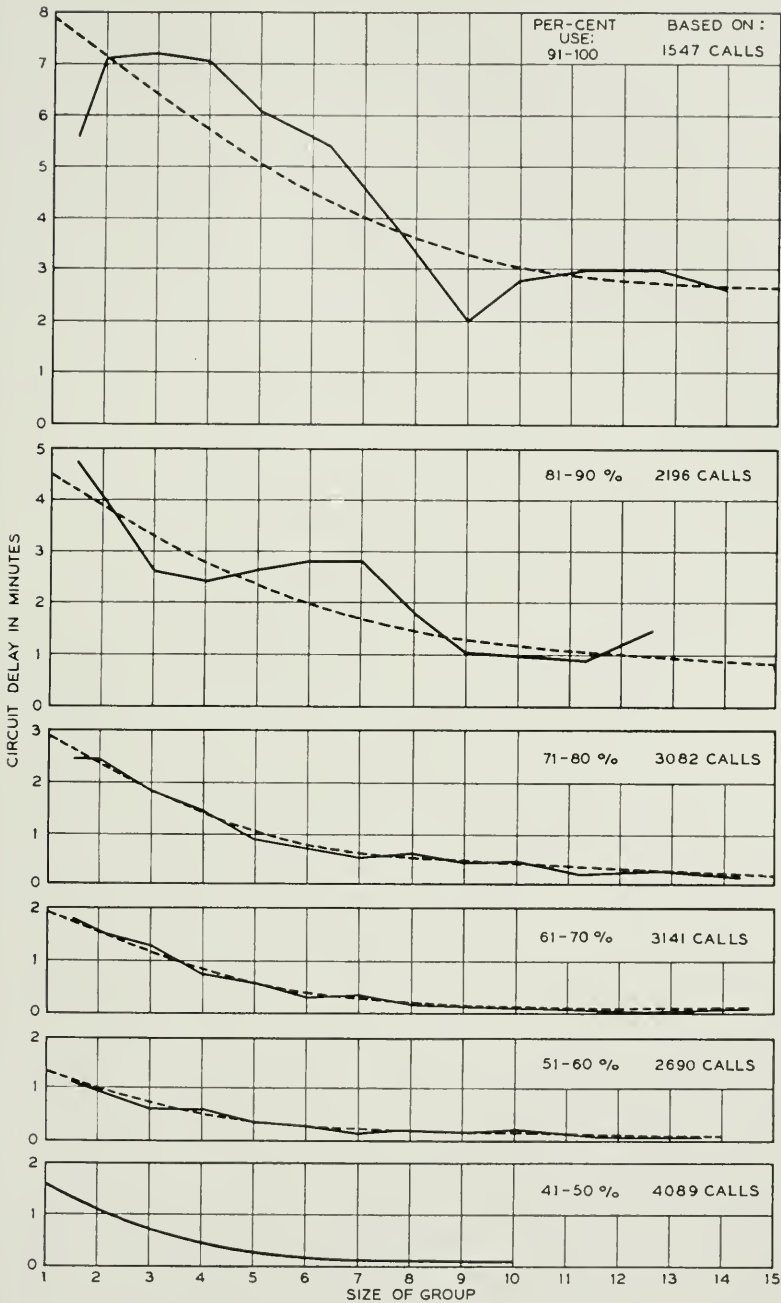
Fig. 1—Average circuit delay on all calls (with alternate routes where authorized).
Circuit delay = average speed on all calls minus average speed on calls which did not
encounter NC.   Based on 3-point moving average.

However, further reference to mathematical studies of telephone traffic indicated that a delay theory based on constant holding times, first developed
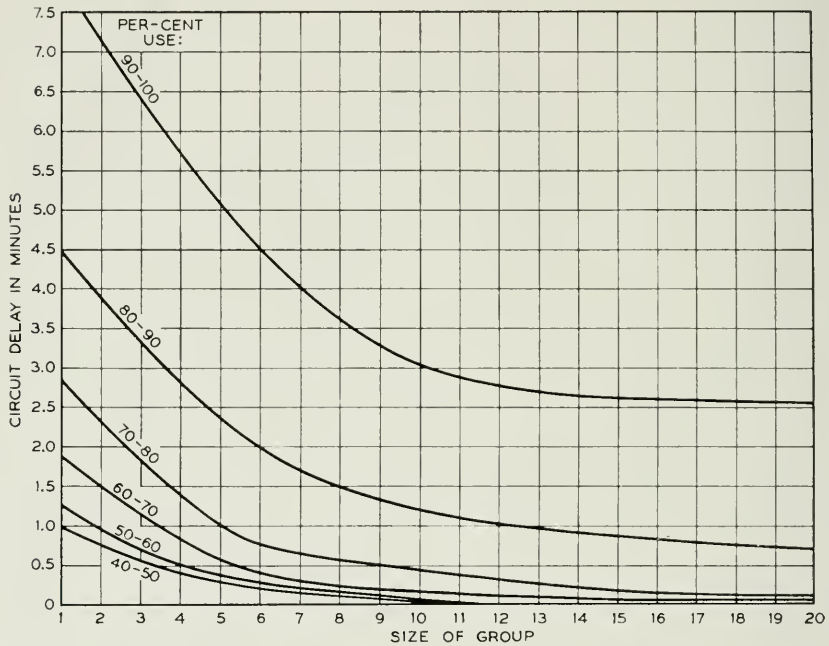
Fig. 2—Average circuit delay on all calls (with alternate routes where authorized). Circuit delay = average speed on all calls minus average speed on calls which did not encounter NC.   Combined curves based on 16,745 calls.

by Felix Pollaczek in Germany and amplified by C. D. Crommelin in England, closely approximated the empirical data. This formula[1] is:

$$d = \sum_{w=1}^{\infty} e^{-aw} \left[\sum_{u=wc}^{\infty} \frac{(aw)^u}{\lfloor u} - \frac{c}{a} \sum_{u=wc+1}^{\infty} \frac{(aw)^u}{\lfloor u} \right]$$

In which

d = average delay on all calls

a = average simultaneous calls submitted to a group of c trunks (trunk hours)

c = number of trunks in group

It may be quite reasonable that the Pollaczek constant holding time formula should better represent toll delays than an exponential holding time formula since the toll charge and perhaps other factors ordinarily cause these calls

[1] C. D. Crommelin, "Delay Probability Formulae," *P.O.E.E. Journal*, Jan. 1934, p. 266

to exhibit considerably less percentage deviation from their average than is found in the exponential distribution.

In order to compare the empirical data with the Pollaczek formula it was necessary to assume a holding time per attempt since the formula expresses the delays in terms of the average interval of use whenever the circuit is in use. The average holding time as reported by the companies for the groups included in the study was 8.3 minutes per message. Recent data show 1.42 attempts per call disposed of. Relating this figure to the 8.3 minutes results in an average holding time per circuit use of 5.85 minutes. Six
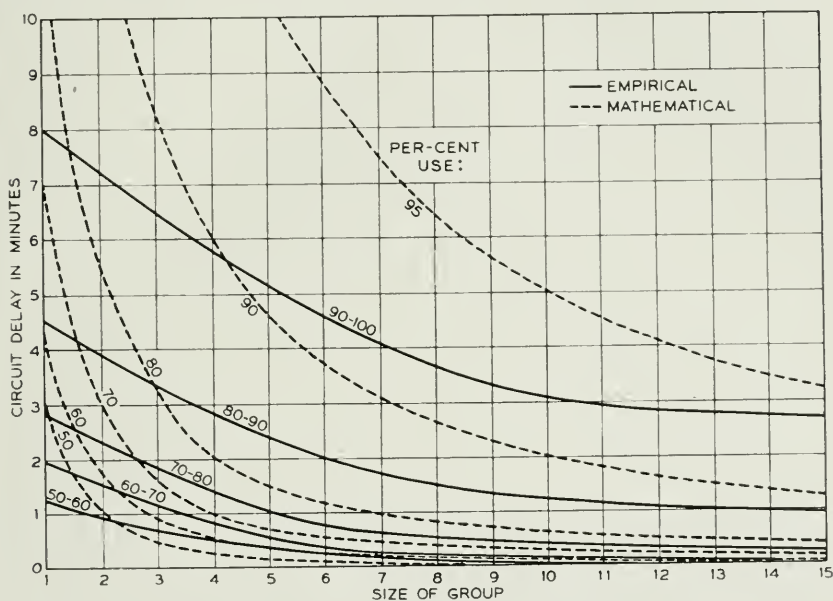
Fig. 3—Average circuit delay on all calls.
Comparison of empirical data (with alternate routes) and Pollaczek formula (no alternate routes) using a 6 minute holding time.

minutes is therefore well within the limits of accuracy required for this purpose.

Curves were prepared from the Pollaczek formula for various levels of usage at a 6-minute holding time per attempt. The corresponding curves derived from the empirical data were then superimposed for comparative purposes as shown in Fig. 3. It will be seen that the shape and levels of the curves are very similar *except* for the smaller groups on which the effect of alternate routes tends to reduce the average length of delay.

As a further check on the validity of the Pollaczek formula, the delay data from the Cleveland (1929–1930) study were expressed in terms of hold-

ing times for different group sizes at different levels of usage and compared with delay intervals developed from the formula. This comparison is

COMPARISON OF 1945 STUDY WITH CLEVELAND STUDY OF 1929–30
Based on 3.5 Min. HT per Circuit Attempt
or 5.25 Min. HT per Message

No. of Trunks	% Use	Minutes Delay		% of H.T.	
		Cleveland	1945	Cleveland	1945
1	55–60	.8	.8	21	22
1	65–70	1.1	1.2	30	35
1	75–80	1.5	1.9	43	53
1	85–90	2.7	3.0	76	86
3	55–60	.3	.5	09	13
3	65–70	.5	.8	14	22
3	75–80	.9	1.3	24	36
3	85–90	1.8	2.4	51	68
6	55–60	.1	.1	03	04
6	65–70	.3	.3	09	08
6	75–80	.6	.6	16	16
6	85–90	1.3	1.7	36	48
10	55–60	.1	—	02	01
10	65–70	.3	.1	07	03
10	75–80	.4	.3	12	08
10	85–90	1.0	1.0	28	28
14	55–60	.1		01	
14	65–70	.2	.1	05	02
14	75–80	.3	.2	09	05
14	85–90	.8	.7	23	19

The minutes of circuit delay shown above for the Cleveland study are derived by subtracting the minimum speed of 1.65 minutes from the actual overall speed for the various sizes of groups and levels of usage. The comparable 1945 figures are taken from Fig. 5.

It will be noted that the principal differences occur on the smaller groups at the higher levels of usage. This is undoubtedly due to the fact that the alternate routes are more heavily loaded today than they were in 1929–30 and therefore are less helpful in absorbing the overflow from the first route.

The holding time used in this comparison as probably typical of 1929–30 is derived as follows:

Conversation time. . . . . 3.00 minutes
Operating time. . . . . 2.25 minutes
———
5.25 minutes

No. of Circuit Uses per Message 1.50
HT per Circuit Use (5.25 ÷ 1.50) 3.50

Fig. 4

shown in Fig. 4. It will be seen that there is substantial agreement between the two sets of delay factors, such differences as there are being explained in the notes on that figure.

The delay in securing a circuit varies directly with the holding time, i.e., a call waiting for a circuit will be delayed twice as long if the group is handling 10-minute calls as would be the case with 5-minute calls. This is best illustrated by one call awaiting access to a single circuit group. The new call may appear at any time during the progress of the existing call, but the average delay for many such delayed calls will be one-half the holding time of the existing call. If the existing call uses the circuit for five minutes, the new call will wait $5 \div 2 = 2.5$ minutes. If the existing call uses the circuit for ten minutes, the new call will wait five minutes. It should be noted that the average delay depends upon the average length of time that the circuit is in use each time that it is used, in other words, the holding time per circuit attempt.

The Cleveland study did not go into this phase in detail, the statement being made that the effect of holding time on speed of service "is slight." This is so when considering the *overall* speed of service, with which that study was primarily concerned, because of the weight of operating and subscriber time intervals. Reference to that study shows that the circuit delay increased about in proportion to the holding time when a minimum operating and subscriber time interval is subtracted from the overall speed, as follows:

	Holding Time		
	5′	7.5′	10′
Total Overall Speed..............................	2.2	2.6	3.0
Minimum Operating and Subscriber Time Interval.................................	1.6	1.6	1.6
Average NC Delay-Trunk Speed................	.6	1.0	1.4

COMBINATION OF MATHEMATICAL AND EMPIRICAL METHODS

Because of the apparent close agreement between the Pollaczek delay formula and two representative large samples of actual NC delay data taken at different periods and under widely different conditions, 1930 and 1945, the Pollaczek formula can be used for deriving expressions of the average duration of NC with intertoll trunk operation without alternate routes. The effect of alternate routes in reducing the duration of NC can be shown with sufficient accuracy for practical needs from the empirical data. The curves shown in Fig. 5 were constructed on this basis. For large groups the formula has been extended in Fig. 6.

It is advantageous to have available an acceptable mathematical formula for expressing the relationship between the loads carried by the trunk groups and the length of time that the average call will be delayed because of NC conditions, i.e., the part played by trunk provision in the overall speed of service. With such a formula results can be predicted for any given set of

assumptions without going through the burdensome process of accumulating actual data in reliable volume and with useful frequency. The formula
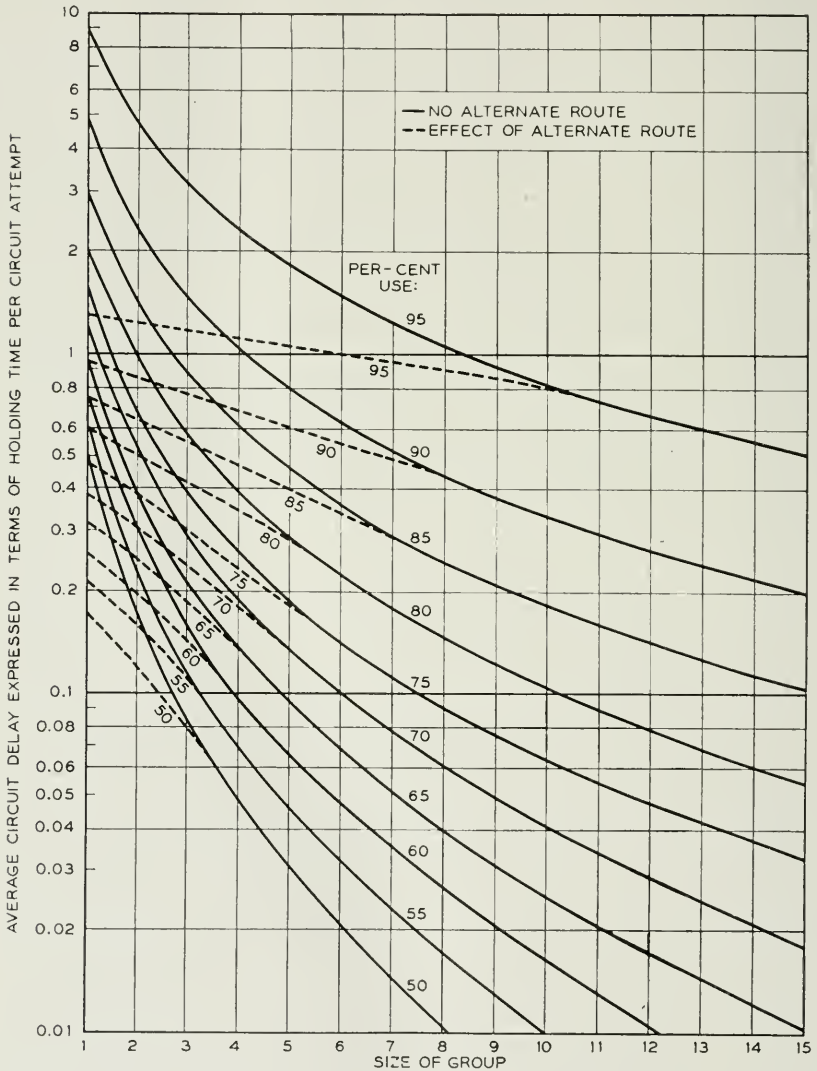
Fig. 5—Average circuit delay on all calls expressed in terms of the holding time per circuit attempt.

also provides a convenient means of checking the adequacy of the trunk facilities in any unusual situations which may be observed from time to time.
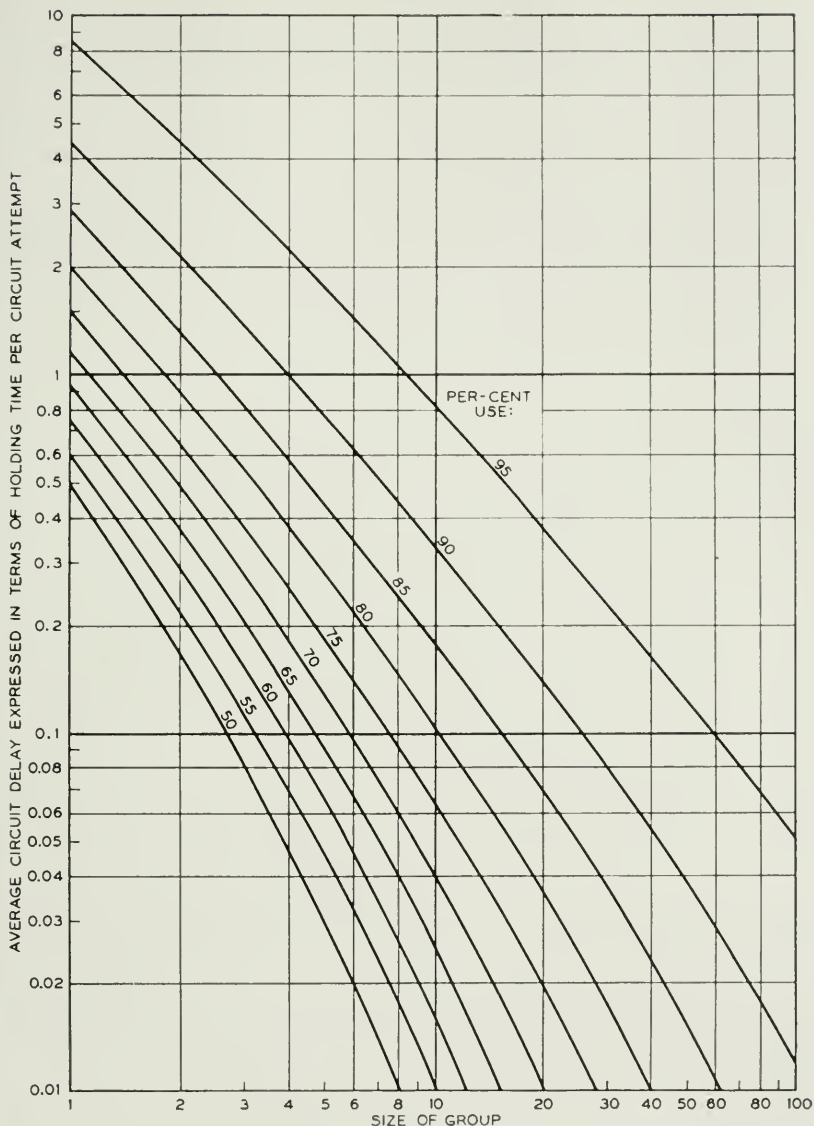
Fig. 6—Average circuit delay on all calls expressed in terms of the holding time per circuit attempt.
Based on Pollaczek formula.

In the case of the alternate route effect, where a variable is introduced which the formula does not encompass, it may be necessary later to recheck

this factor by observed data should the conditions governing the use of
alternate routes change substantially or should the actual results at some
future time on groups provided with alternates be found to differ from those
currently predicted.

Delays are expressed in Figs. 5 and 6 as a percentage of the holding time
per *average use* of the trunk.   The holding time factors used in intertoll
trunk engineering generally are expressed in terms of the holding time
per message.   Therefore, in order to use the curve conveniently it is neces-
sary to reduce the holding time per message to the holding time per trunk
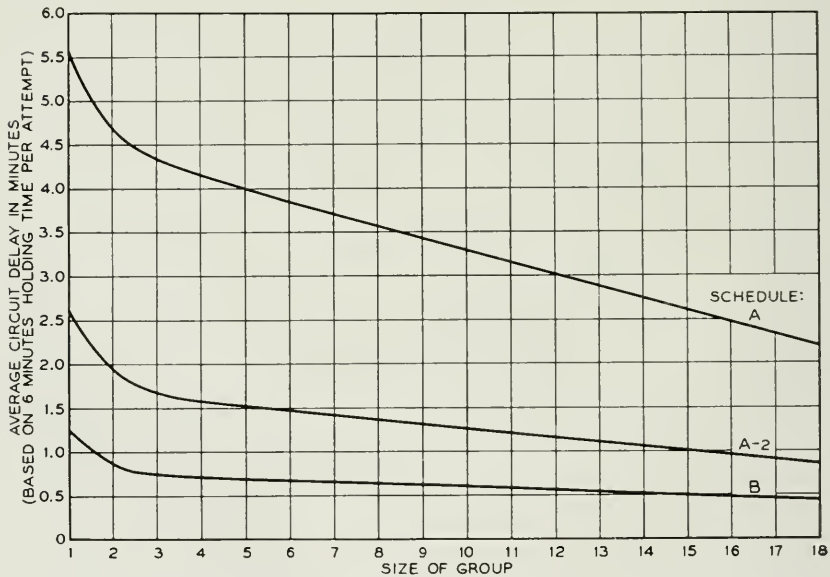
Fig. 7—Average circuit delay on all calls resulting from present intertoll trunk capacity
schedules (without alternate routes)—Using Pollaczek formula.

use or attempt.   This can be done with sufficient accuracy for this purpose
by a ratio of 1.5 outward attempts per call disposed of.

DESCRIPTION OF ENGINEERING TABLES

Having determined the average circuit delay, as previously described, the
delays which result from the present Intertoll Trunk Capacity Tables A,
A2, and B can be determined by referring the percentages of use on these
tables to the curves in Figs. 5 and 6.   Figure 7 indicates that each of these
existing tables results in variable delays depending upon the size of the
group.

The curves in Figs. 5 and 6 can also be used to construct new capacity tables which should produce a relatively uniform delay for any size of group. If we select an average delay of three-tenths of a holding time as an example
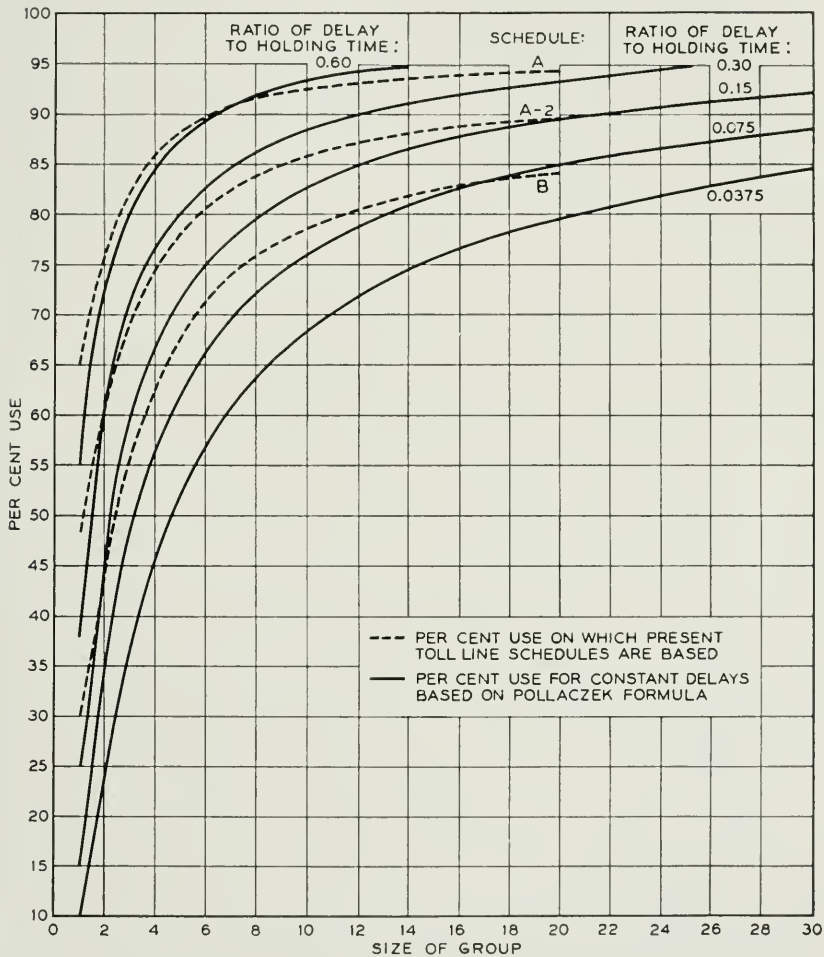
Fig. 8—Per cent. use with present intertoll trunk schedules and for constant delays based on Pollaczek formula.

and follow the .3 line across these curves, we see that a two-trunk group can be kept busy 60.8 per cent. of the time; a three-trunk group can be in use about 71.1 per cent. of the time; about 76.7 per cent. for four trunks, etc.

Figure 8 shows the per cent. usage for groups of from one to thirty trunks

which result in average delays of .0375, .0750, .15, .30, and .60 of a holding time. The usage obtained from present capacity tables is also shown for comparison. From this figure it will be seen that five new tables based on these average delays will adequately cover the field encompassed by those now in use. The usage curves for these five selected delay intervals were
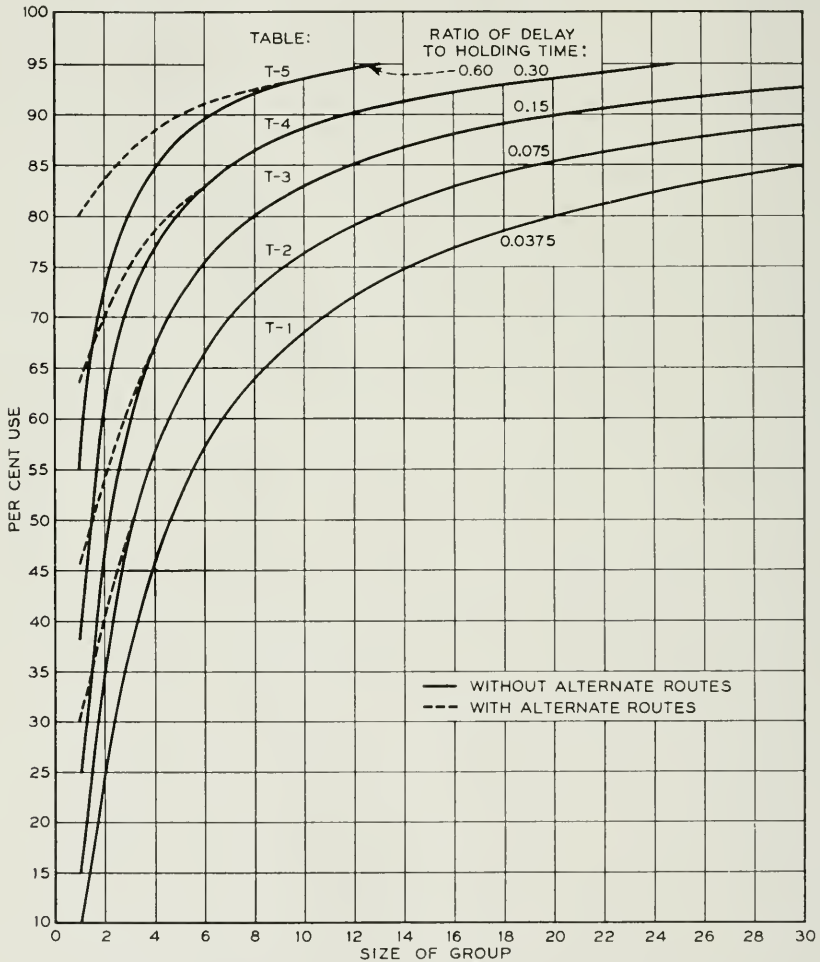
Fig. 9—Per cent. use for constant delays based on Pollaczek formula.

redrawn in Fig. 9, on which the effect of alternate routes is also shown. Capacity Tables T-1 to T-5 (Fig. 10) were constructed from this curve, Table T-1 representing the shortest (.0375) delay and Table T-5 the longest (.60) delay.

For situations where it may be necessary for service reasons to provide trunks on a basis more liberal than Table T-1, the interlocal trunk capacity

tables should be used. These latter tables, constructed from the Poisson formula, express a service relationship in terms of the percentage of calls *encountering NC* rather than the average *duration of NC*. When trunks are provided on a basis as liberal as that implied by the interlocal tables the frequency of NC is deemed to be the more important service consideration because the duration of NC is so short as to be of lesser consequence.

Since the Pollaczek formula expresses the delay as a percentage of the holding time per *circuit attempt* but the intertoll trunk engineer is accustomed to dealing with holding times per *message* the delays indicated for each of Tables T-1 to T-5 have been expressed in the latter terms, using a 1.5 ratio of attempts per message. This is consistent with previous computations. With Table T-4 and a five-minute message holding time, for example, we have:

5-minute HT per message $\div$ 1.5 = 3.33 minutes HT per attempt

3.33 minutes x .30 delay = 1.0 minute delay

This one minute delay shown in the heading of Table T-4 is, therefore, actually the delay per attempt when the holding time per message is five minutes. The delay per attempt is the important criterion because the "speed of toll service" as quoted from service observations is generally the speed of the first attempt and the two are, therefore, comparable.[2]

As the usage approaches 100 per cent. there may be an indeterminate backed-up potential demand and the normal relationship between service and loading no longer holds true. From other data assembled for this purpose, it appears that about 96–97 per cent. represent the practical upper limit of usage, beyond which trunk speeds can not be accurately predicted. For practical reasons, therefore, group capacities have not been computed for percentages of use higher than 97 even though, from a theoretical viewpoint, the curves derived from the Pollaczek formula would permit extending the usage virtually to 100 per cent. This would also apply to Tables T-1 and T-2 if they were extended above 75 trunks.

RELATION OF CIRCUIT DELAYS IN BUSY HOUR TO TOTAL DAY

Up to this point the trunk speed of service has been discussed in terms of the busiest hour. However, the overall speed of service is generally quoted in terms of the total day. Therefore, one additional step is necessary, namely to determine the relationship of the trunk speed in the busy hour to that of the total day.

---

[2] There is one exception to the statement that the "speed of toll service" is the speed on the first attempt. That is the case of a built-up connection where an NC condition is encountered at an intermediate office which persists so long that the first circuit is released. When the connection is established it is at least the second attempt. The full time interval during which the ticket is held awaiting completion is included in the speed quoted on that call. Similar intervals were also included in the empirical data for this study and the results are, therefore, comparable in this case also.

Fig. 10—Intertoll trunk capacity tables

No. of Trunks	Table T-1 (.13, .20, .28 Minutes) No Alt. Route % Use	Min-utes Capacity	Alt. Route % Use	Min-utes Capacity	Table T-2 (.25, .40, .55 Minutes) No Alt. Route % Use	Min-utes Capacity	Alt. Route % Use	Min-utes Capacity	Table T-3 (.5, .8, 1.1 Minutes) No Alt. Route % Use	Min-utes Capacity	Alt. Route % Use	Min-utes Capacity	Table T-4 (1.0, 1.6, 2.2 Minutes) No Alt. Route % Use	Min-utes Capacity	Alt. Route % Use	Min-utes Capacity	Table T-5 (2.0, 3.2, 4.4 Minutes) No Alt. Route % Use	Min-utes Capacity	Alt. Route % Use	Min-utes Capacity
1	10.0	6	—	—	15.0	9	30.0	18	25.0	15	50.0	30	38.3	23	63.4	38	55.0	33	80.0	48
2	25.0	30	—	—	35.0	42	40.0	48	48.3	58	53.3	64	60.8	73	70.0	84	73.3	88	83.4	100
3	37.2	67			48.9	88			60.0	108	61.1	110	71.1	128	75.0	135	80.6	145	86.6	156
4	46.3	111			56.7	136			67.5	162			76.7	184	78.3	188	85.0	204	88.4	212
5	52.7	158			62.3	187			72.3	217			80.3	241	80.7	242	87.6	263	90.0	270
6	57.5	207			66.7	240			75.4	272			83.1	299			89.5	322	91.1	328
7	61.2	257			70.0	294			78.1	328			85.0	357			91.0	382	91.9	386
8	64.2	308			72.7	349			80.2	385			86.5	415			92.1	442	92.5	444
9	66.7	360			74.8	404			81.9	442			87.6	473			93.0	502		
10	68.8	413			76.5	459			83.1	499			88.7	532			93.6	561		
11	70.6	466			78.0	515			84.2	556			89.6	591			94.1	621		
12	72.2	520			79.3	571			85.3	614			90.3	650			94.5	680		
13	73.6	574			80.4	627			86.2	672			90.9	709			94.9	740		
14	74.8	628			81.4	684			86.9	730			91.4	768			95.2	800		
15	75.9	683			82.3	741			87.6	788			91.9	827			95.5	860		
16	76.9	738			83.1	798			88.1	846			92.3	886			95.8	926		
17	77.8	793			83.9	855			88.6	904			92.6	945			96.1	980		
18	78.6	848			84.5	912			89.0	962			93.0	1,005			96.3	1,040		
19	79.3	904			85.0	969			89.5	1,021			93.4	1,065			96.5	1,100		
20	80.0	960			85.5	1,026			90.0	1,080			93.7	1,125			96.6	1,160		

Trunk Speed when average H.T. per Msg. is: 4-6 Mins., 7-9 Mins., 10-12 Mins.

21	80.7	1,017	86.0	1,084	90.4	1,139	94.0	1,185	96.8	1,220
22	81.3	1,074	86.5	1,142	90.8	1,198	94.3	1,245	97.0	1,280
23	82.0	1,131	87.0	1,200	91.1	1,257	94.5	1,305		
24	82.5	1,188	87.4	1,258	91.4	1,316	94.8	1,365		
25	83.0	1,245	87.7	1,316	91.7	1,375	95.0	1,425		
30	85.0	1,530	89.7	1,606	92.8	1,670	95.9	1,725		
35	86.5	1,816	90.5	1,901	93.7	1,908	96.4	2,025		
40	87.8	2,106	91.5	2,196	94.6	2,268	96.8	2,325		
45	88.8	2,398	92.3	2,491	95.1	2,568				
50	89.7	2,692	92.9	2,786	95.6	2,868				
55	90.5	2,987	93.4	3,081	96.0	3,168				
60	91.2	3,282	93.8	3,376	96.3	3,468				
65	91.7	3,577	94.3	3,675	96.6	3,768				
70	92.2	3,872	94.7	3,975	96.8	4,068				
75	92.7	4,170	95.0	4,275	97.0	4,368				

Note:—The Trunk Speeds of Service indicated at the heading of each Table represent the average busy hour delay in securing a trunk on each attempt, when the holding times per message are as shown.

To develop this relationship it was necessary first to determine a typical distribution of traffic throughout the day, i.e., the ratio of the busy hour to each of the other hours. Actual delays experienced on a particular group may deviate somewhat from those developed herein to the extent that the actual distribution varies from the typical distribution.

The probable total day circuit delays are derived from Figs. 5 and 6 and from a typical distribution of traffic based on a five-day record on each of 20 groups in Ohio and 24 groups in Illinois.

Hours	No. of Calls	% of Total Traffic in 14-hour day (Used as weighting factor)
1 (Busy Hour)	100	12.80
1	90	11.53
1	90	11.53
1	80	10.25
1	80	10.25
1	75	9.62
1	70	8.98
1	65	8.35
1	55	7.07
5	75	9.62
14	780	100.0
10	20	
24	800	

It will be noted that the above distribution shows a busy hour which is 12.5% of the total 24-hour day but that the weighting factors are based on a total day of 14 hours. This corresponds to the normal service observing period so that the results will be comparable with the overall total day speeds obtained from service observations.

The total day delays for Tables T-1 to T-5 for each size of group were computed as illustrated in the following sample calculation:

TABLE T-5—FIVE TRUNKS

Hours	% of BH	% Use in BH (Table T-5)	% Use Each Hour	Circuit Delay (Fig. 6)	Weight Factor	Weighted Delay
1	at 100	× 87.6 =	87.6	.60 ×	12.80% =	.0768
1	90		78.8	.26	11.53	.0300
1	90		78.8	.26	11.53	.0300
1	80		70.2	.13	10.25	.0133
1	80		70.2	.13	10.25	.0133
1	75		65.7	.10	9.62	.0096
1	70		61.3	.07	8.98	.0063
1	65		56.9	.06	8.35	.0050
1	55		48.2	.03	7.07	.0021
5				None	9.62	.0000
14					100.00	.1864

The figure .1864 derived above represents the average delay expressed as a per cent. of the holding time per circuit use (attempt). The last step, therefore, is to relate this figure to the message holding times contemplated in Table T-5, as follows:

Holding Time Per Message		Ratio of Attempts Per Message		Holding Time Per Attempt		% of H. T. Delayed		Average Delay
5 min.	÷	1.50	=	3.33	×	.1864	=	.62 min.
8 min.	÷	1.50	=	5.33	×	.1864	=	1.00 min.
11 min.	÷	1.50	=	7.33	×	.1864	=	1.37 min.

The results of similar calculations are summarized in Fig. 11.

As pointed out previously, actual delays experienced will deviate from those shown in Fig. 11 to the extent that the actual hourly distribution varies from that which has been used. If a particular group has a higher per cent busy hour the total day delays should be less than indicated. Conversely, if the group has a lower per cent busy hour the delays should be greater. However, the variations in distribution which are most likely to be encountered in practice will not have any marked effect on the total day delays except possibly for groups of about five trunks or less which are loaded as heavily as indicated in Tables T-4 and T-5.

Per Cent. NC Encountered

The per cent. of calls delayed by NC as noted by the operators on the tickets analyzed for this study was plotted for each level of usage in steps of 10% as shown in Fig. 12, using a 3 point moving average. Each of these curves was then redrawn in relation to the others and the combined results are in Fig. 13. The results are very similar to those obtained in the Cleveland study of 1929–30, as shown in the same figure.

It should be noted that there is a difference between the per cent. calls encountering NC and the per cent. NC existing. In the present study no data were obtained to indicate NC existing. However, the Cleveland study included such data which showed that the NC existing follows the Erlang "B" formula (Fig. 14) in this respect. The individual points in Fig. 14 were derived by selecting from the Erlang "B" table of overflows the point at which the call-seconds carried (offered minus overflow) gave the desired level of usage.

The difference between NC existing and NC encountered may be due to several factors, some of which are suggested below:

1. Effect of alternate routes.

	Table T-1			Table T-2			Table T-3			Table T-4			Table T-5		
	Holding Time—In Minutes														
	5	8	11	5	8	11	5	8	11	5	8	11	5	8	11
	Trunk Speed—Busy Hour—In Minutes														
No. of Trunks	.13	.20	.28	.25	.40	.55	.5	.8	1.1	1.0	1.6	2.2	2.0	3.2	4.4
	Trunk Speed—Total Day—In Minutes														
1	.09	.14	.20	.19	.31	.43	.38	.60	.83	.67	1.07	1.47	1.20	1.92	2.64
2	.08	.13	.18	.16	.25	.34	.29	.47	.65	.52	.83	1.14	.92	1.48	2.03
3	.07	.11	.16	.14	.22	.30	.25	.39	.54	.43	.69	.95	.77	1.24	1.70
4	.06	.10	.14	.12	.19	.26	.22	.35	.48	.39	.62	.85	.68	1.09	1.50
5	.06	.09	.13	.11	.18	.24	.20	.32	.44	.35	.56	.77	.62	1.00	1.37
6	.05	.09	.12	.10	.17	.23	.18	.29	.40	.33	.52	.72	.57	.91	1.25
7	.05	.08	.11	.10	.15	.21	.17	.28	.38	.31	.49	.67	.53	.85	1.17
8	.05	.08	.11	.09	.14	.20	.16	.26	.36	.29	.46	.63	.50	.80	1.10
9	.05	.07	.10	.09	.14	.19	.16	.25	.34	.27	.44	.60	.48	.76	1.05
10	.04	.07	.10	.08	.13	.18	.15	.24	.33	.26	.42	.58	.46	.73	1.00
11	.04	.07	.09	.08	.13	.18	.14	.23	.32	.25	.41	.56	.44	.70	.96
12	.04	.07	.09	.08	.12	.17	.14	.22	.31	.24	.39	.54	.43	.68	.94
13	.04	.06	.09	.08	.12	.17	.13	.21	.29	.23	.37	.52	.41	.66	.91
14	.04	.06	.09	.07	.12	.16	.13	.21	.29	.23	.36	.50	.40	.64	.88
15	.04	.06	.08	.07	.11	.15	.13	.20	.28	.22	.35	.48	.39	.63	.86
16	.04	.06	.08	.07	.11	.15	.12	.20	.27	.22	.35	.48	.38	.61	.84
17	.04	.06	.08	.07	.11	.15	.12	.19	.26	.21	.34	.47	.38	.60	.83
18	.04	.06	.08	.07	.11	.15	.12	.19	.26	.21	.33	.45	.37	.59	.81
19	.03	.05	.07	.06	.10	.14	.11	.18	.25	.20	.33	.45	.37	.59	.81
20	.03	.05	.07	.06	.10	.14	.11	.18	.24	.20	.32	.44	.36	.58	.79
21	.03	.05	.07	.06	.10	.14	.11	.18	.24	.20	.32	.43	.35	.57	.78
22	.03	.05	.07	.06	.10	.13	.11	.17	.24	.19	.31	.43	.35	.56	.77
23	.03	.05	.07	.06	.09	.13	.11	.17	.23	.19	.31	.42	.35	.55	.76
24	.03	.05	.07	.06	.09	.13	.10	.17	.23	.19	.31	.42	.34	.55	.76
25	.03	.05	.07	.06	.09	.13	.10	.17	.23	.19	.30	.41	.34	.54	.75
30	.03	.05	.06	.05	.09	.12	.10	.15	.21	.18	.28	.39	.33	.52	.72
35	.03	.04	.06	.05	.08	.11	.09	.15	.21	.17	.27	.37	.32	.51	.70
40	.03	.04	.06	.05	.08	.10	.09	.14	.20	.16	.26	.36	.31	.50	.68
45	.03	.04	.06	.05	.07	.10	.09	.14	.19	.16	.26	.35	.30	.49	.67
50	.02	.04	.05	.05	.07	.10	.09	.14	.19	.16	.25	.34	.30	.48	.66
55	.02	.04	.05	.04	.07	.10	.08	.14	.19						
60	.02	.04	.05	.04	.07	.10	.08	.13	.18						
65	.02	.04	.05	.04	.07	.09	.08	.13	.18						
70	.02	.04	.05	.04	.07	.09	.08	.13	.18						
75	.02	.04	.05	.04	.06	.09	.08	.13	.18						

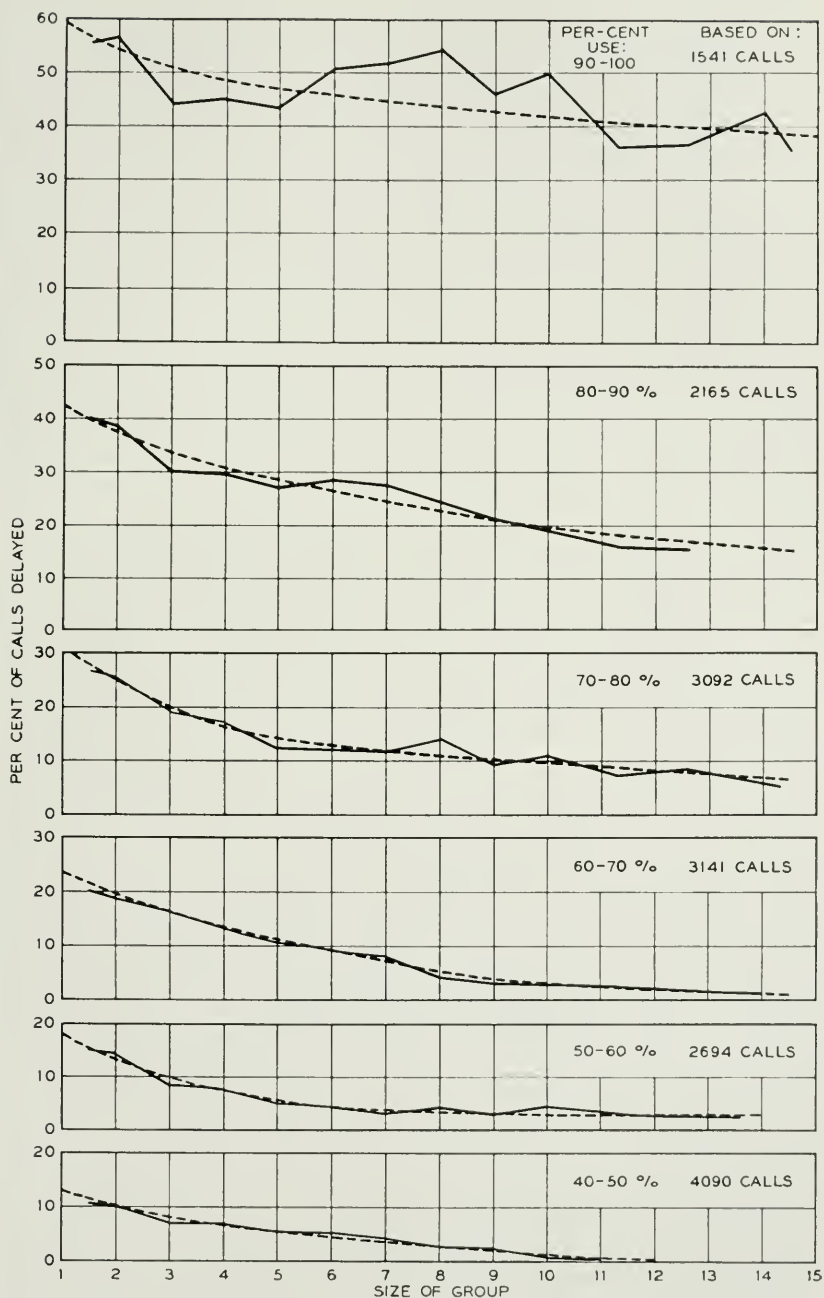Fig. 11—Relation of trunks speeds in busy hour to total day

Fig. 12—Per cent. of calls delayed by NC as noted by the operators (with alternate routes where authorized).

2. The operators did not make NC notations on the tickets in a certain proportion of the cases where NC was actually encountered.

3. Because the operator does not test for an idle trunk with machine-like finality there are probably many cases where she does not consider that an NC condition existed if it was of such short duration that it did not materially affect her ability to secure a trunk.

4. The possibility of some "limited sources" effect in the case of small groups. The number of people in Newark who have occasion to call York, Pa. must be relatively small since only one trunk is provided. Therefore, while the trunk is in use on one call there is less likelihood that a second call will be originated than would be the case if there were a greater community of interest between the two places. Although the NC condition exists during the period of one call, no NC is encountered unless there is a second and overlapping call.
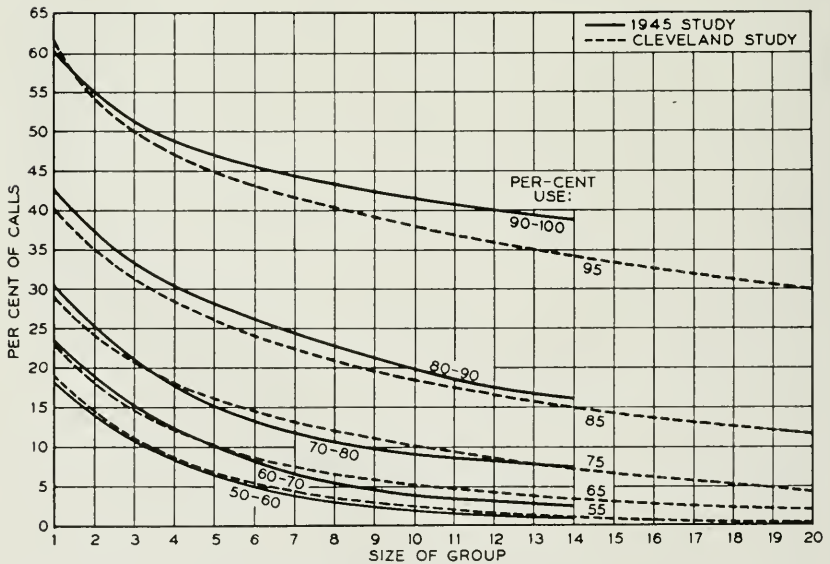

Fig. 13—Per cent. of calls delayed by NC (with alternate routes where authorized).

Because of the importance of Items 2 and 3 above, both of which involve the testing of trunks by operators, the empirical data should be used as representative of per cent. NC encountered with ringdown operation (operator testing) and the Erlang "B" per cent. NC existing should be used as representative of intertoll dialing conditions (mechanical testing).
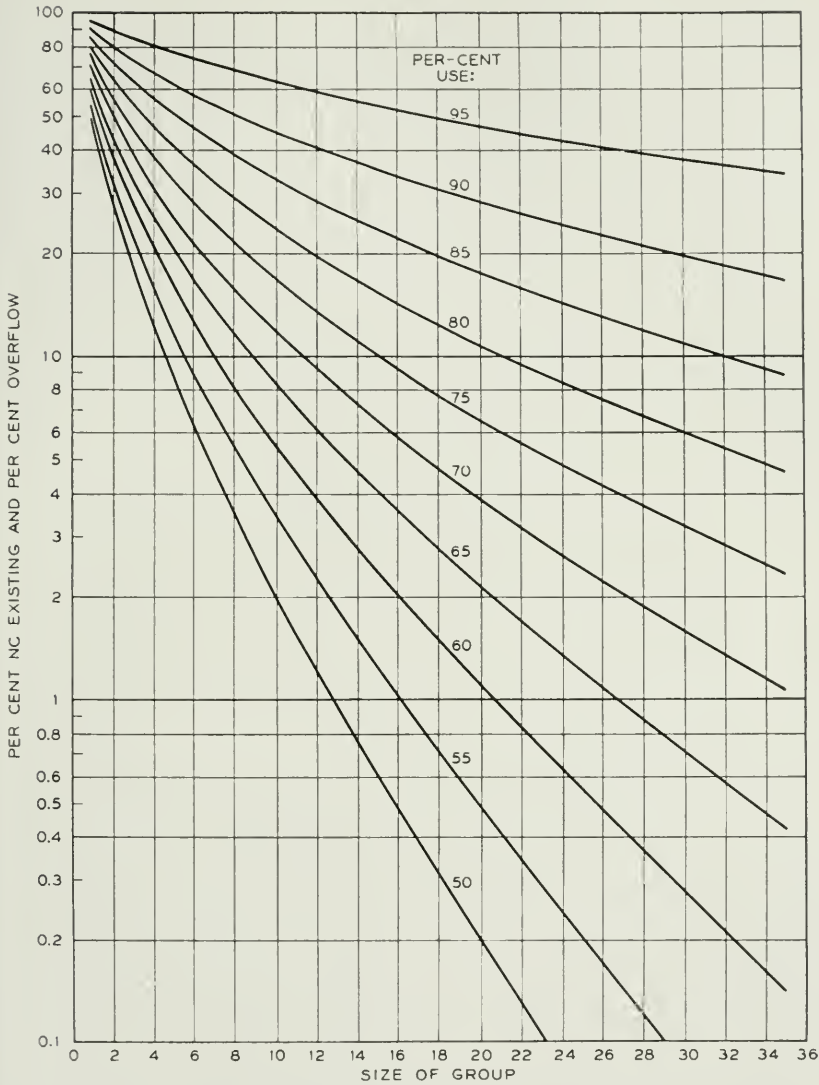
Fig. 14—Per cent. NC existing and per cent. overflow at various levels of usage. Based on Erlang "B" formula.

It is interesting to note that each of the proposed Capacity Tables T-1 to T-5 results in a fairly uniform percentage of NC encountered by the operators as determined from the empirical data (Fig. 13) and also a fairly uniform

percentage of NC existing as determined from the Erlang "B" formula (Fig. 14).   Using Table T-1 as an example, the NC conditions are as follows:

% Use	No. of Trunks Table T-1 from Fig. 9	%NC Encountered from Fig. 13	% NC Existing from Fig. 14
50	4.6	6.5	10.0
55	5.5	6.0	10.0
60	6.7	6.0	10.7
65	8.4	6.5	10.9
70	10.7	7.0	11.0
75	14.1	7.5	11.0
80	20.0	7.0	10.7
85	30.0		10.8

Similar comparisons made with the other capacity tables indicate similar uniformity in the most frequently used portion of the tables, i.e., up to 20 or 30 trunks.   The results are as follows:

Capacity Table	% NC Encountered by Operators (from empirical data) With Alternate Routes for the Small Groups	% NC Existing (from Erlang "B") Without Alternate Routes
T-1	6–7	10–11
T-2	9–11	18–19
T-3	14–18	27–29
T-4	21–26	39–42
T-5	33–35	55–57

Conclusion

Since the primary function of an intertoll trunk capacity table is to translate a desired speed of toll service into the number of trunks required for that level of service, the table used should be indicative, within reasonable limits, of the probable effect of trunk provision on the overall speed. For this reason, tables which reflect a uniform service situation will be more useful in intertoll trunk engineering and administration than the present tables which have inherent service variations.   Capacity tables such as Tables T-1 to T-5 will therefore be substituted for present Schedules A, A2 and B.

Spark Gap Switches for Radar

By F. S. GOUCHER, J. R. HAYNES, W. A. DEPP and E. J. RYDER

INTRODUCTION

A N ESSENTIAL feature of radar is the generation, by means of an oscillator, of high-energy pulses of short duration, repeated many times a second. The energy for these pulses is furnished to the oscillator from a power supply in a variety of ways. One of the most widely used of these is the "line type modulator" in which a pulse-forming network made up of a series of condensers and inductances is charged from the power supply through a choke and is then discharged by a switch so that a substantially constant current will flow for a predetermined short time through the primary of a pulse transformer coupled to the oscillator. This switch is, therefore, an essential component of this type of modulator.

To meet the pulsing requirements of radar as it developed during the war, this line modulator switch was required to withstand thousands of volts between pulses and to carry hundreds of amperes for the pulse duration which was of the order of microseconds. Also, the switching operation had to be repeated from a few hundred to a few thousand times a second for a total operating time of hundreds of hours. Furthermore, the dissipation of energy within the switch had to be very small in comparison with the energy delivered to the oscillator for efficient operation.

The switch which had the widest application in this type of modulator was that employing an electric spark. Of over 50,000 radars of various types manufactured by the Western Electric Co. during the war, over half employed the electric spark in switching. One form of this switch was a rotary spark gap, operating in air, in which the timing of breakdown was controlled mechanically. These gaps were successfully adapted to a variety of radar types including airborne radar. However the demands for a more compact and lighter weight switch capable of operating at lower voltages for airborne radar led to the development of fixed sealed unit type gaps which, when connected in series, can be broken down electrically in a simple circuit.

Many problems had to be solved in the development of these switches. They required a considerable amount of study, and with the aid of new techniques developed during the war, a number of significant measurements have been made which have extended our knowledge of sparks generally. It is the object of this paper to describe the results of some of these studies,

563

as well as to describe the essential characteristics of a variety of spark gap switches which were used in such numbers that they may be considered as an important contribution to the war effort.

I. Rotary Spark Gap Switches For Low Voltage Circuits

Rotary gaps were used successfully as switches in some of the earlier radar systems developed by Bell Telephone Laboratories.   The switching voltages in the modulator circuits were relatively high, being in excess of 20 kilovolts. No trouble was encountered in switching at the required pulsing rates nor in obtaining satisfactorily long life.   Fortunately the sparks tend to move about the electrode surfaces uniformly and the rate of erosion is such that with tungsten or molybdenum electrodes a uniformly small change in electrode dimensions is achieved which in no way interferes with satisfactory operation over long periods of time.

A difficulty was encountered, however, when the switching voltage was reduced to lower values, as required for applications in which the power supply voltages were limited.   The gaps failed to break down regularly.

A particular application in which this difficulty was encountered was one in which the power supply was limited to 4 kilovolts, and in which 80 ampere pulses of one microsecond duration were required every 600 microseconds.   The modulator circuit used was that shown schematically in Fig. 1 (a).   The pulse-forming network includes the condenser elements which are charged through the choke and discharged by the spark gap designed to break down at the required pulsing rate of 1600 per second.   The load is the primary of a pulse transformer coupled to a magnetron and is closely equivalent to a 50-ohm resistance.   The constants of the circuit are such that following the discharge of the network it is recharged sinusoidally along the solid line of Fig. 1 (b) to a peak value of approximately 8000 volts in 600×10^{-6} seconds, at which point breakdown must again occur and the operation be repeated.   The dashed line is the approximate path of the charging voltage wave when breakdown at the peak fails to occur.

A rotary spark gap was designed to meet these pulsing conditions.   In this gap there are four fixed and four moving electrodes as indicated in Fig. 1 (a).   These electrodes are tungsten rods 3 mm in diameter and about 15 mm in length mounted with their axes parallel and so spaced that the moving electrodes pass very close to the fixed electrodes with an overlap of about one-half their length.   The speed of the moving electrodes is such that in the region of near approach the maximum gradients are those indicated in Fig. 1 (c).   The solid curve shows the gradients when breakdown takes place at the required time and the dashed curve the gradients when breakdown fails to occur.   Although the latter greatly exceed the normal

dielectric strength of air, sparking failed to take place a large fraction of the time.
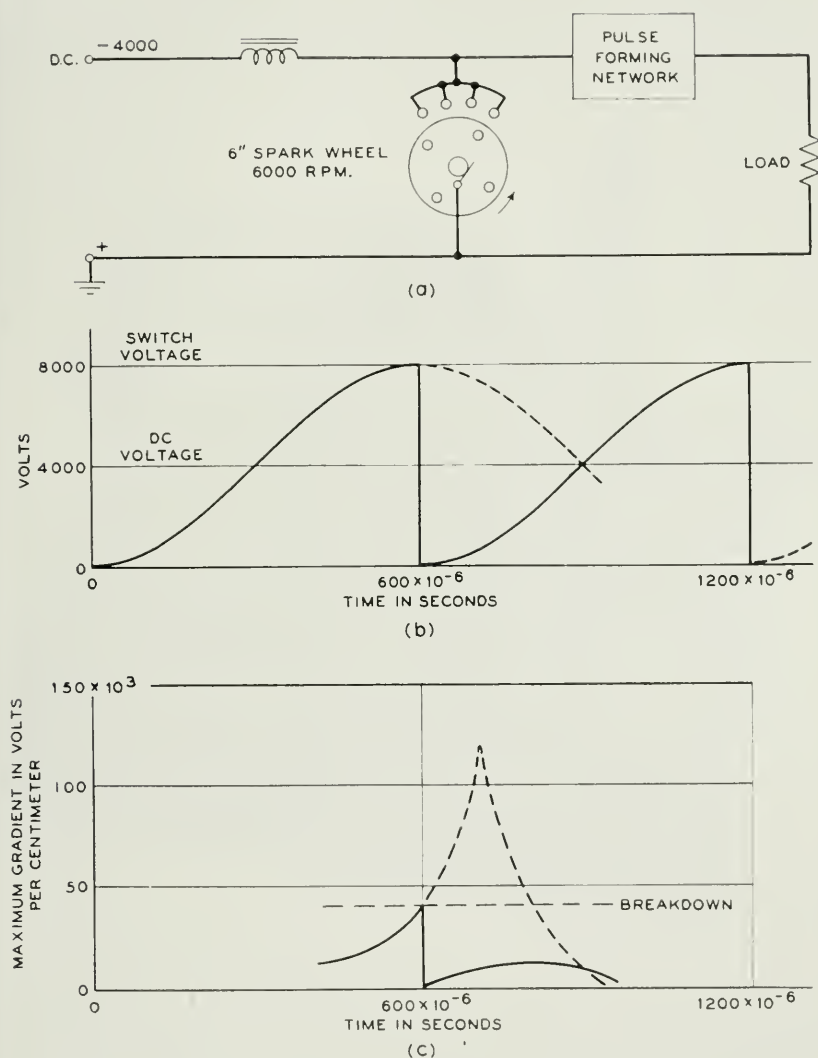

Fig. 1—(a) Line modulator circuit with rotary spark gap switch, (b) switch voltage vs. time, (c) maximum voltage gradient between electrodes vs. time.

Experiment indicated that this was caused by spark delay time, as irradiation of the cathodes by means of an ultra-violet lamp produced 100%

breakdown. This method of reducing spark delay time was not practical, however, and other means were sought. A solution was found through a rediscovery of the efficacy of corona prior to breakdown which came about through the introduction of a properly placed sharp edge on the cathode. Although this edge was apart from the sparking area of the cathode, 100% breakdown of the gaps was obtained and spark delay time so reduced
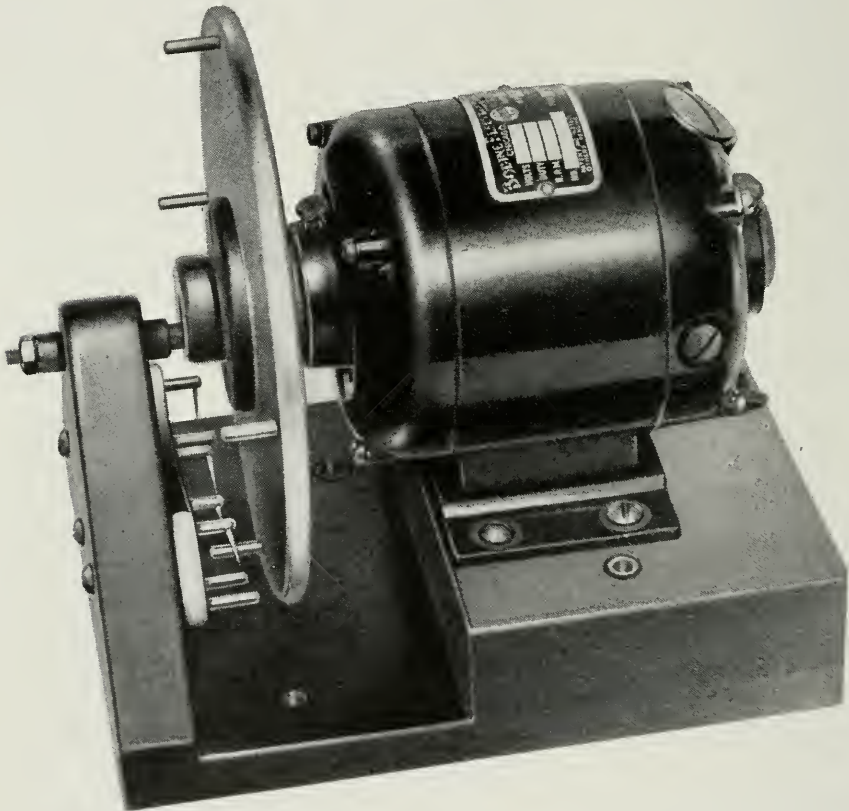

Fig. 2—Experimental model of rotary gap showing corona points.

that mechanical limitations alone controlled the variation in time of breakdown.

The essential features of the gap as finally developed for this project are shown in the photograph of an experimental model, Fig. 2, and in the perspective drawing and accompanying diagram, Fig. 3. The electrodes are of tungsten as in the earlier design, and corona points are introduced by the addition of the rods holding sharp metal points mounted on the same metal

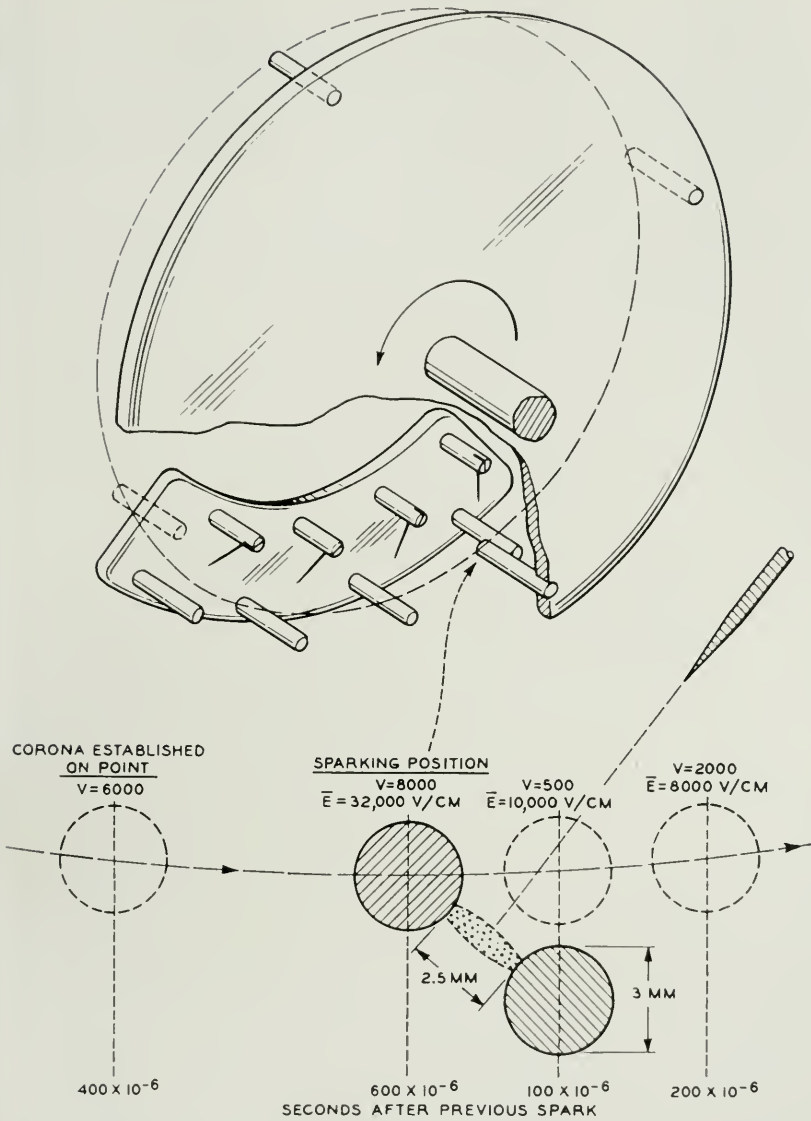Fig. 3—Above, perspective drawing of rotary gap showing arrangement of corona points. Below, diagram showing voltage (*V*), and mean voltage gradient (*E*) at various times in the spark-over region.

base as the fixed electrodes. The moving electrodes pass between the fixed electrodes and their associated corona points. This arrangement is clarified in the diagram which is a section through a plane normal to the electrode

axes and passing through the region of overlap. The shaded areas are for the sparking position as indicated and the location of the corona point is shown to scale. Experiment shows that when the moving electrode has reached the position corresponding to 400×10^{-6} seconds after the previous spark, corona is established on the point. Thus the cathode is irradiated for 200×10^{-6} seconds prior to breakdown.

No serious erosion problem was encountered when these gaps were operated for many hundreds of hours in air. No deterioration of the points was observed when their locations were properly adjusted so as to avoid sparking over to them. The cathode erosion rate is so low that appreciable flats were produced only after a hundred hours of operation. The anode erosion was estimated to be less than one-tenth of that of the cathode, and was doubtless associated with a small amount of reverse current shown to be present. The magnitude of the cathode erosion rate for tungsten in air is about twenty-five fold less than that for tungsten in hydrogen under the same conditions which indicates that oxygen plays an important and somewhat unexpected role in making practical the operation of these gaps.

There was, however, a serious corrosion problem when these gaps were adapted to airborne radar because of the necessity for sealing the modulator unit in a container capable of maintaining atmospheric pressure at high altitudes. Spark discharges in air are attended by the formation of both ozone and oxides of nitrogen, the latter combining with moisture to form nitrous and nitric acids. These reached such concentrations under continuous operation in the container that they were damaging to all enclosed equipment because of their corrosive action. A solution for this was arrived at after considerable study on the part of the Chemical Department. This consisted of the use of a copper impregnated activated carbon as an absorbent. With this absorbent a life of 500 hours was shown to be possible.

Over 10,000 rotary gap switches of this type were manufactured and used successfully in both ship and airborne radars. However, under the urge to reduce the weight of all possible components used in airborne radar and even to eliminate the necessity for pressurizing, the development of glass-enclosed fixed gaps as switches was diligently pursued.

The authors would like to acknowledge the cooperation of Mr. N. I. Hall of the Whippany Laboratories whose responsibility it was to engineer and develop these rotary gap switches for manufacture.

II. Fixed Gaps

Preliminary experiment indicated that a series of fixed gaps could be made to operate satisfactorily as a modulator switch. A study was therefore made to determine the most suitable gas atmosphere, electrode material and gap design for use in sealed gaps. This led to the development of a unit

type gap, two or more of which could be operated in series.   The first unit type gap had an aluminum cathode and a hydrogen-argon gas atmosphere. Later, under the urge for higher peak powers, mercury cathode gaps were developed.   Details of this study and development will be discussed in this section.

(a) *Triggering Gaps in Series*

An alternative to a rotary gap in which the timing of spark breakdown is controlled mechanically was the use of a fixed gap, the breakdown of which is controlled electrically.   One method of accomplishing this was to use a third electrode to which an impulse voltage was applied periodically at double the frequency of the resonant charging circuit.   This voltage breaks down one gap with a discharge of energy furnished by the trigger circuit, which in turn causes a breakdown of the main gap, either through a modification of the field in this gap or through the addition of ions which reduce its breakdown voltage.   This type of gap, however, required a strong air blast to de-ionize the gaps and, because of this, its use obviously presented no great improvement over the rotary gap.   It was well known that the rate of de-ionization is greater the smaller the gap, so an attempt was made to trigger without air blast a number of smaller gaps which when connected in series would withstand the full switch voltage as employed in the rotary gap.

The arrangement used was that shown in Fig. 4.   Six tungsten pins, 3 mm in diameter, were mounted with their axes parallel and spaced to give five 0.5 mm gaps.   The switch voltage was divided by means of equal high resistances connected across the gaps, and a highly damped bi-directional trigger pulse was applied to the four middle pins through capacity coupling as shown.   Corona points were also connected in such a way that the cathode of each of the gaps is irradiated in order to reduce the spark delay time.

By an appropriate adjustment of the circuit elements it was demonstrated that this series of gaps could be broken down by the trigger pulse and de-ionized with sufficient rapidity so that no air blast was required.

Although no attempt will be made here to elucidate the detailed steps in the triggering of the five gaps just described, we can get a qualitative idea of the process by considering a simple two-gap and three-gap circuit which, it turned out, was all that was required for the various applications of fixed gaps as they were eventually developed.

In the two-gap circuit, Fig. 5 (a), if the first half cycle of the trigger pulse and the switch voltage are both positive, gap 1 will break down when the potential at the mid point, due to the sum of the switch voltage and that of the trigger, is equal to the gap breakdown voltage, which for the moment we shall consider as singly valued.   This effectively shorts gap 1 and throws the full switch voltage across gap 2 which in turn will break down provided this

switch voltage is equal to or greater than the breakdown voltage of one gap.
The gaps will operate for all switch voltages up to a value equal to twice the
breakdown voltage of one gap when both gaps will break down without the
addition of trigger.   This, then, is the maximum operating voltage and the
ratio of maximum to minimum operating voltage is two to one on the basis
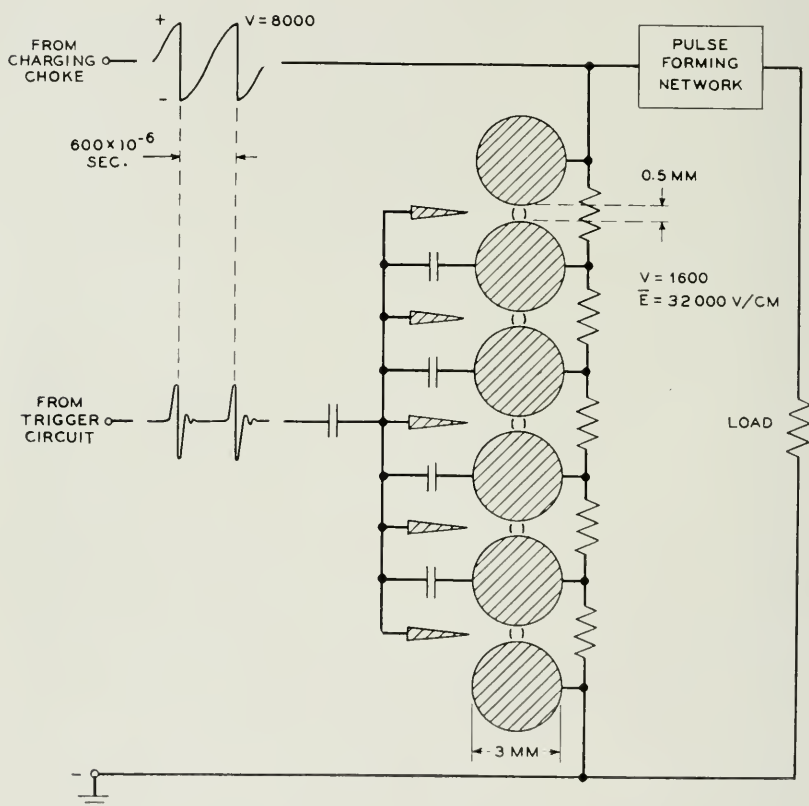of this simple picture.

Fig. 4—Line modulator circuit with fixed gap switch composed of five 0.5 mm. air gaps
triggered electrically.

In the case of the three-gap circuit, Fig. 5 (b), gaps 1 and 2 may be broken
down by the simultaneous application of a trigger pulse through capacity
coupling.   The circuit elements can be so chosen that gap 1 first breaks down
leaving enough trigger on gap 2, over and above that furnished by the switch
voltage, to break it down.   The full switch voltage is then applied across
gap 3 and it will break down for values of switch voltage in excess of the

single gap breakdown voltage. In this case the switch voltage may be increased to a value three times that of the breakdown of one gap before the three gaps can break down without addition of trigger. Thus the ratio of maximum to minimum operating voltage is three to one. Ideally this ratio may be increased by the addition of more gaps.
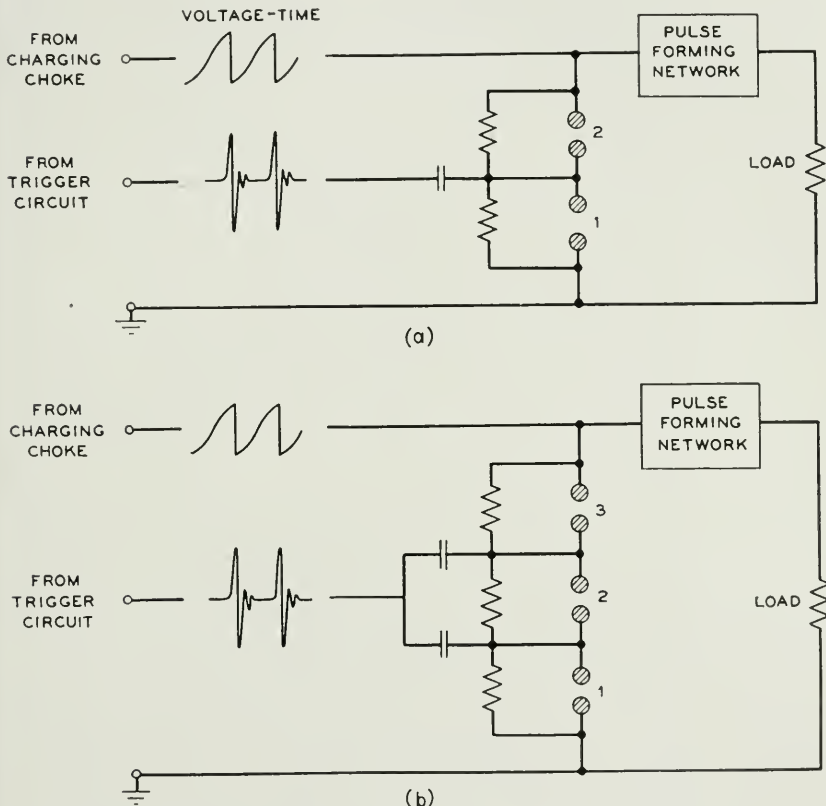
Fig. 5—Line modulator circuit (a) using two fixed gaps as switch, (b) using three fixed gaps as switch.

The operating characteristics of actual gaps do not conform exactly to this simple picture as we shall see later. This is because the breakdown voltage of a gap is not singly valued but depends on a variety of conditions such as rate of rise of applied voltage, pulsing rate, and the energy of the pulse, as well as the type of gap employed. However, we may regard it as a qualitatively correct picture of the operating characteristics of series gaps.

A more complete description of operating characteristics will be given in a

later section, but, in view of the fact that the gap itself plays an important part in these characteristics, it seems desirable to describe first the gap types with which we have to deal.

(b) *The Hydrogen-Argon Aluminum Cathode Gap*

Following the successful triggering of fixed gaps in air without the use of air blast for their de-ionization, experiments were undertaken with sealed gaps in various gas atmospheres using simple rod electrodes having their axes parallel. A large number of gases were tested and the conclusion reached that hydrogen was the most satisfactory because of its high de-ionization rate. With it fewer and wider gaps were required to meet a given pulsing condition. Three 4 mm. gaps in hydrogen at pressures somewhat less than atmospheric were approximately equivalent to the five 0.5 mm. gaps in air already referred to. Thus, from this point of view, the use of hydrogen would very greatly simplify the problem of making practical gaps.

The spark in hydrogen, particularly with relatively small peak currents, was, however, unsatisfactory in that it terminated in a high-pressure glow with a high cathode drop rather than the low drop required for efficient switching. The addition of about 25% argon corrected this and about this proportion was used successfully in the gaps with which we are concerned in this report.

Although the required operating conditions were met with this gas mixture, cathode erosion or sputtering was so excessive with all readily available cathode materials that this factor appeared as the chief obstacle in the way of making practical gaps. The sputtered material was deposited on all surfaces in the form of a fine powder which eventually destroyed the insulation, thereby limiting the useful life of the gaps to a few hours.[1]

A promising lead was, however, obtained in the case of aluminum cathodes. It was observed that some of the sputtered material deposited on the anodes opposite the cathodes from which it was removed. This deposit was reasonably compact and smooth, which suggested the possibility of reducing by gap design the extent of harmful scattering. This might be achieved by increasing the amount of sputtered cathode material which is deposited on the anode or returned to the cathode within the sparking area.

The tube, Fig. 6, was an early attempt in this direction. This tube had three 4 mm. gaps between flat electrodes, the cathode surfaces having raised portions to confine the sparking within their areas. The gaps were

---

[1] At about this time we learned that the British had developed sealed gaps triggered by means of an auxiliary electrode and known as "Trigatrons." These were high pressure gaps containing argon with a small amount of oxygen to reduce sputtering of the electrodes. The life of these gaps was determined by the time required to clean up this oxygen. Though these were tried it was decided to follow an independent development avoiding if possible all clean up effects.

operated successfully for somewhat over 100 hours before enough scattered material accumulated to interfere with gap insulation. A uniform spark distribution was maintained throughout this time and measurement showed that aluminum was removed quite uniformly from the raised portion of the cathode to a depth of only a fraction of a millimeter. An equally thick
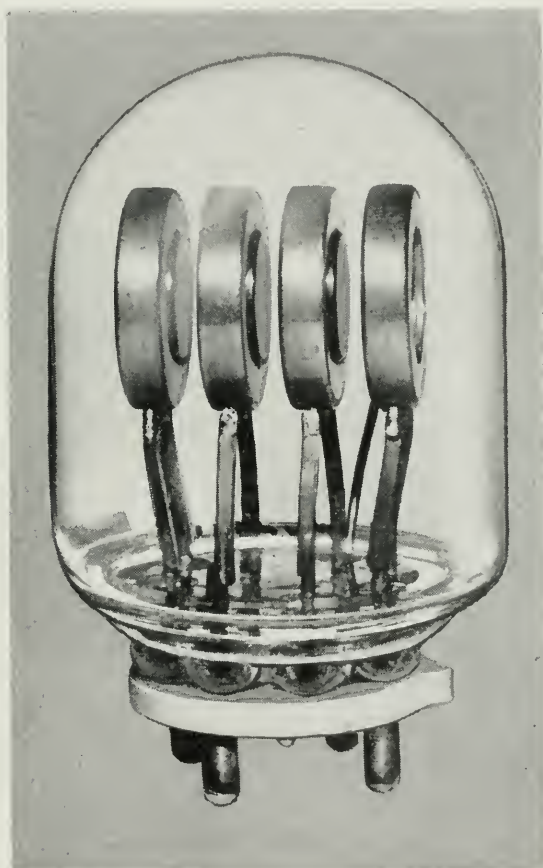
Fig. 6—Three gap tube having aluminum electrodes and a hydrogen-argon atmosphere ——actual size.

though somewhat rougher deposit was formed on the opposing anode surface, thereby retaining the gap spacing very satisfactorily. About 30 milligrams of loose material were scattered throughout the tube.

A more drastic but also more successful design change was introduced by making three separately enclosed gaps, one of which is shown in the photo-

graph and radiograph, Fig. 7. In these gaps the sparking area of the cathode was hemispherical in shape, partly surrounding a spherical anode. These gaps were operating successfully at the end of 1000 hours. The scattered material was deposited on only a portion of the glass envelope of
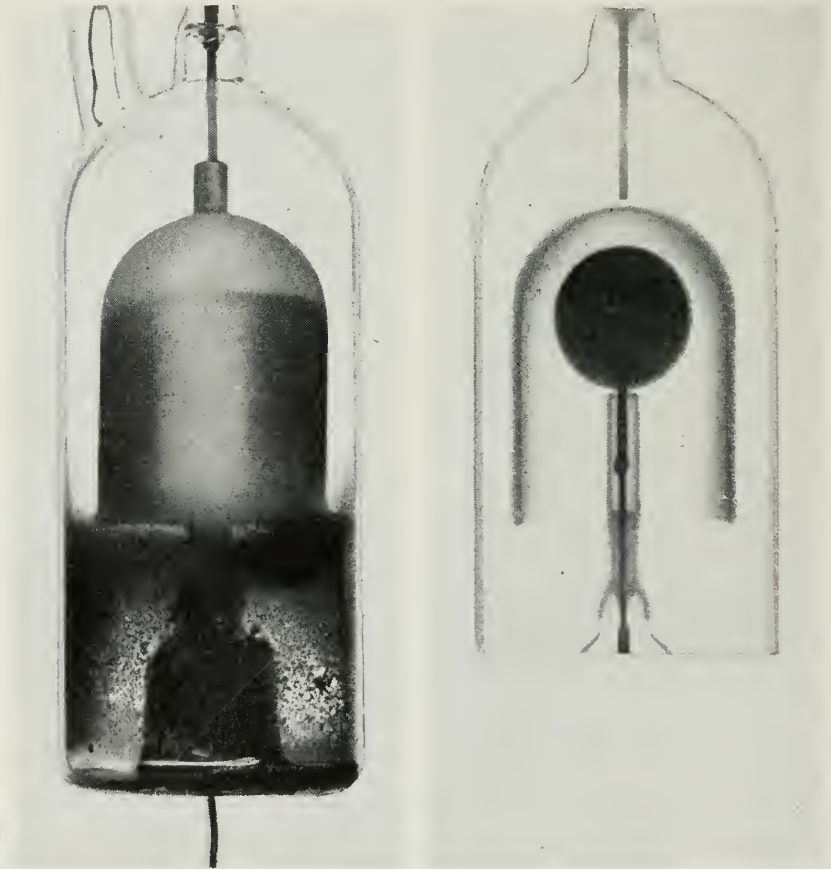
Fig. 7—Photograph and radiograph—actual size—of the first unit type gap having a re-entrant aluminum cathode, a spherical aluminum anode and a hydrogen-argon gas atmosphere, after operating 1000 hours with a 40 ampere pulse of one microsecond duration repeated 1660 times a second.

each gap, as shown in the photograph. The extent of the material removed from the cathode and deposited on the anode, as shown in the radiograph, was such as to cause no marked change in gap spacing. Furthermore, the operating range remained substantially constant throughout the 1000 hours of operation as shown in Fig. 8. This was an important observation since it

indicated that there is no gas clean-up effect associated with gap operation, a fact that was later proved by careful measurement of gas pressure before and after operating gaps of this type. A section through the anode of this gap, Fig. 9 (a), shows that the anode deposit is not compact but assumes the form of a coral-like structure. This low-density deposit must, however, be electrically equivalent to a compact surface as shown by the constancy of the operating characteristics with time.

In view of the success of this design it was decided to develop gaps of the unit type having anodes well enclosed by the cathode surfaces. An attempt to make a more practical gap is that shown in the photograph and radio-
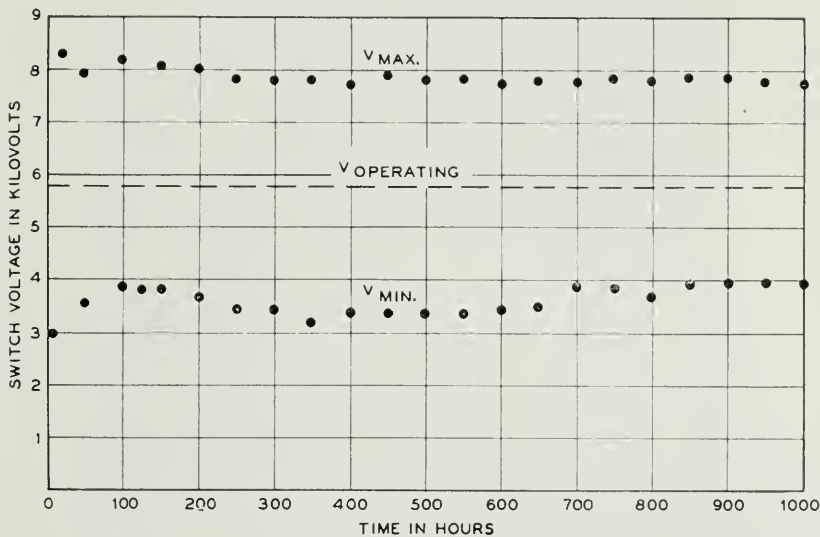
Fig. 8—Maximum and minimum operating voltages as a function of time, for three unit gaps of the type shown in Fig. 7, when operated in series.

graph, Fig. 10, both of which were taken after 750 hours operating time. In this gap the anode is an aluminum rod rounded at the end mounted concentrically with the enclosing cathode which has a hemispherical closed end. The corona point was added to facilitate starting. Because of the higher anode gradient the sparking was confined to the end region of the tube as indicated in the radiograph, and for this reason we have designated this design an "end sparking tube". A section through the anode, Fig. 9 (b), shows a deposit which in this case is compact due to the fact that the moving spark is confined to a smaller area than in the previous tube, Fig. 7 (a). It is to be noted also that the scattered material is less in extent than that

obtained with the first design, pointing to a more effective trapping of the sputtered material.

Weight loss measurements made under a variety of pulsing conditions show that though the rate of cathode erosion is somewhat dependent on the
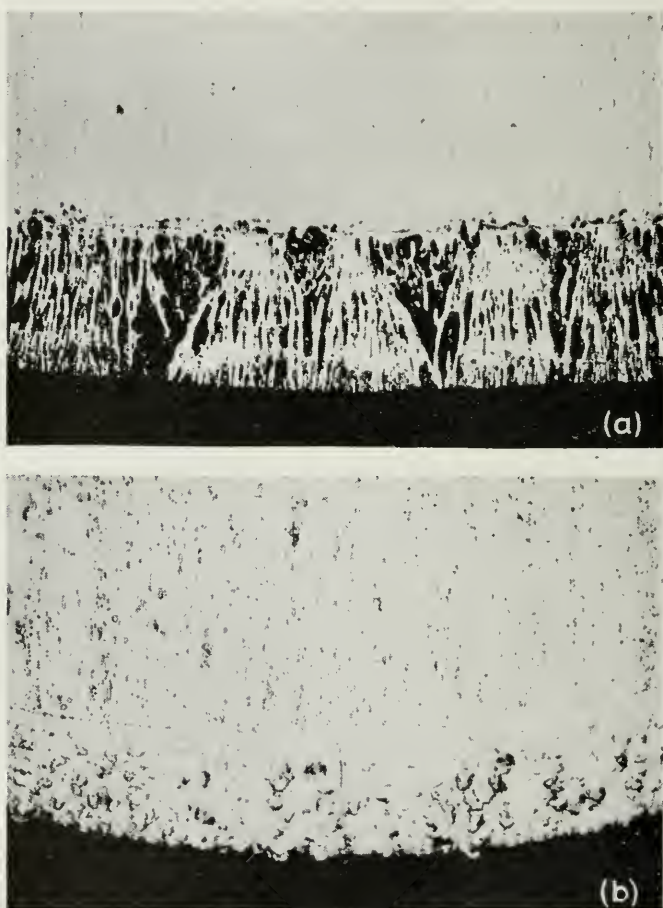
Fig. 9—Photomicrographs (×50) of sections showing anode deposits for (a) unit gap shown in Fig. 7, (b) unit gap shown in Fig. 10.

pulse duration, it depends to a much greater extent on gap design. Erosion rate measurements in terms of grams per coulomb are shown in Fig. 11 for the two gap designs there indicated and for pulse durations varying from one to five microseconds. It is clear that the open gap type of design in which the cathode is small leads to a loss which is at least five fold greater than

that of the "end sparking" type of gap. The smaller loss in the case of the latter shows that much more material returns to the cathode for resputtering

Fig. 10—Photograph and radiograph—actual size—of end sparking unit gap, after operating 700 hours with a 65-ampere pulse of one microsecond duration and repeated 1660 times a second.

than in the case of the former and supports the use of the unit type gap in which this process can be utilized. A cylindrical cathode enclosing a rod

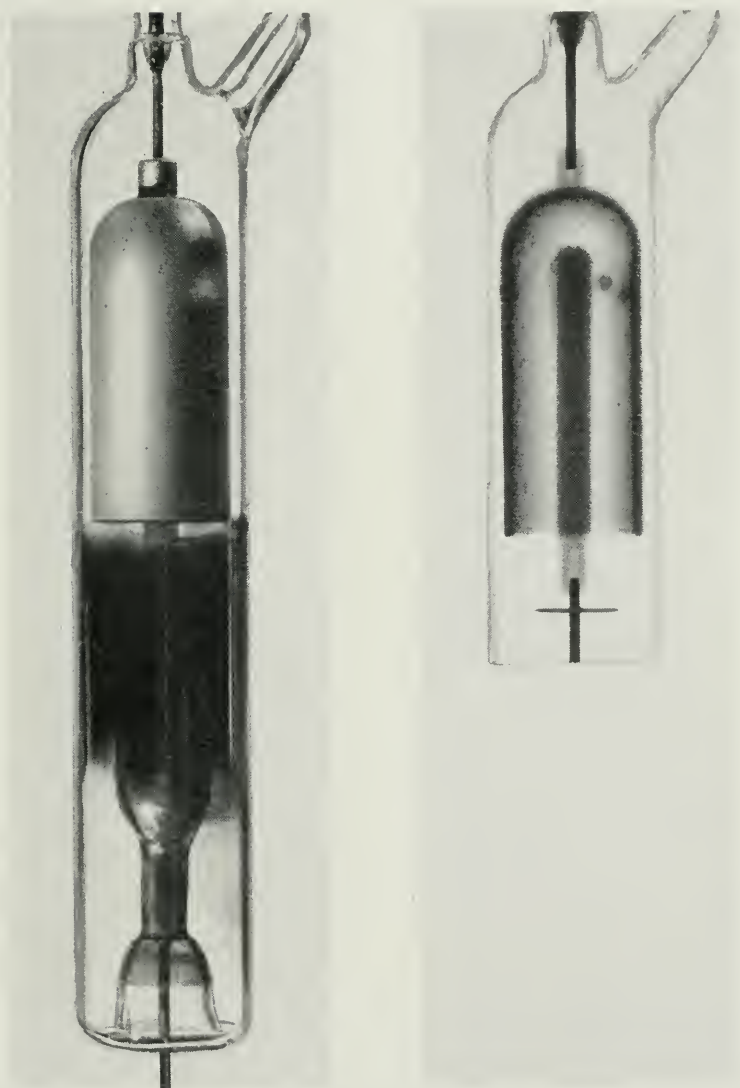anode also behaves in this way and its erosion rate differs but little from the "end sparking" type of tube; in fact, the practical gaps to be described in II-(f) are essentially of this type.

With these facts in mind it would appear that gaps could be designed to meet a variety of pulsing conditions if the total number of ampere hours for a pre-assigned life were known, for the electrode areas could be so adjusted that the changes in gap spacing would be as small as required. Analysis of the gradients associated with the end sparking type of gap shows that there
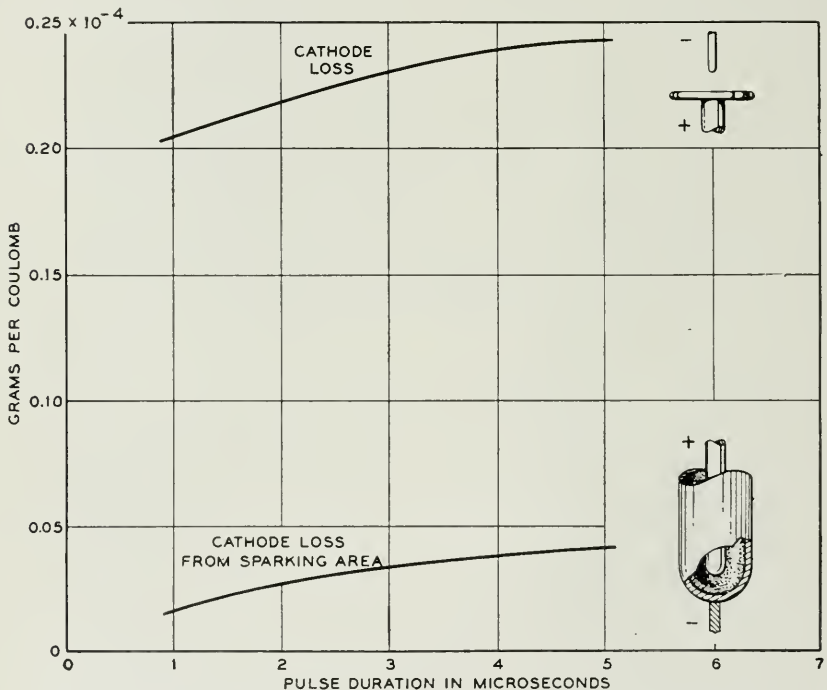
Fig. 11—Cathode loss, in grams per coulomb, as a function of pulse duration showing effect of gap design.

can be a considerable build-up on the anode before there is much change in the maximum gradient which determines the spark-over voltage.

Experience with gaps designed for a variety of pulsing conditions showed that substantial anode build-ups could be tolerated without interfering with operating conditions, but not as much as theory would predict for an unexpected factor had a controlling influence on gap life. This factor was the failure of the spark to keep moving under certain conditions with the result that spikes were grown on the anode which introduced a rapid deterioration

of the operating range due to an increase in anode gradient and also in part to a decrease in gap spacing.

Both the relatively large anode build-up, which may be tolerated without interference with gap operation, and the nature of spike growth, which limited useful life, are illustrated in the radiographs, Fig. 12. It is to be noted that the spike is almost of uniform cross section along its length and radiographs made at various stages of its formation show that growth takes place
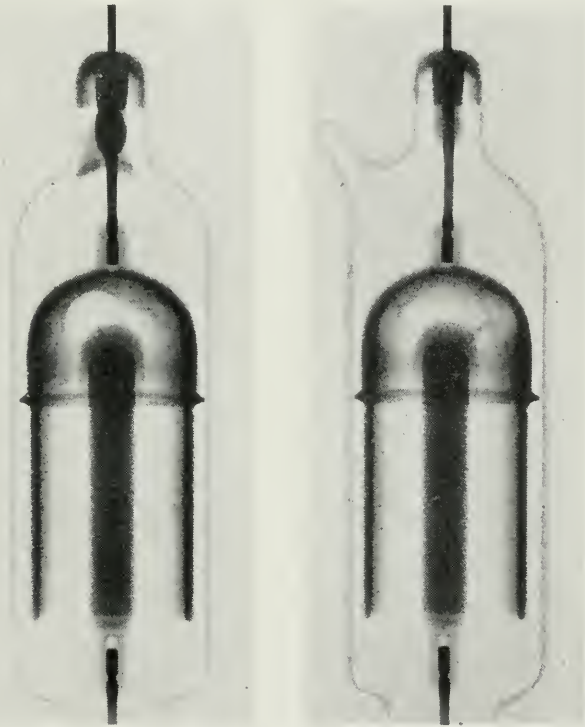
Fig. 12—Radiographs showing two views of the uniform deposit and subsequent spike growth on the anode of an end sparking unit gap.

at its end. This indicates a high concentration of negative ions in the vapor prior to deposit on the anode.

Life test data in which the pulse repetition rate was kept constant at 200 per second are shown in Fig. 13 for gaps having a fixed spacing but in which the peak current is varied (a), and for gaps having a variety of spacings but in which the peak current is kept constant (b). The life is measured in terms of hours to the beginning of spike growth. Both "end sparking" and "side sparking" tubes were employed in the tests. These data clearly show that

if lives longer than 500 are to be obtained, there is a limiting peak current of about 70 amperes with gap spacings of 250 mils or with a peak current of 70 amperes there is a minimum spacing of 250 mils. Similar data were obtained indicating a different critical spacing for other pulsing conditions.

This factor of a critical gap spacing imposed an important restriction on gap design for it was desirable to make gap spacing as small as possible for any given project. This follows because of gap size and weight, also—as
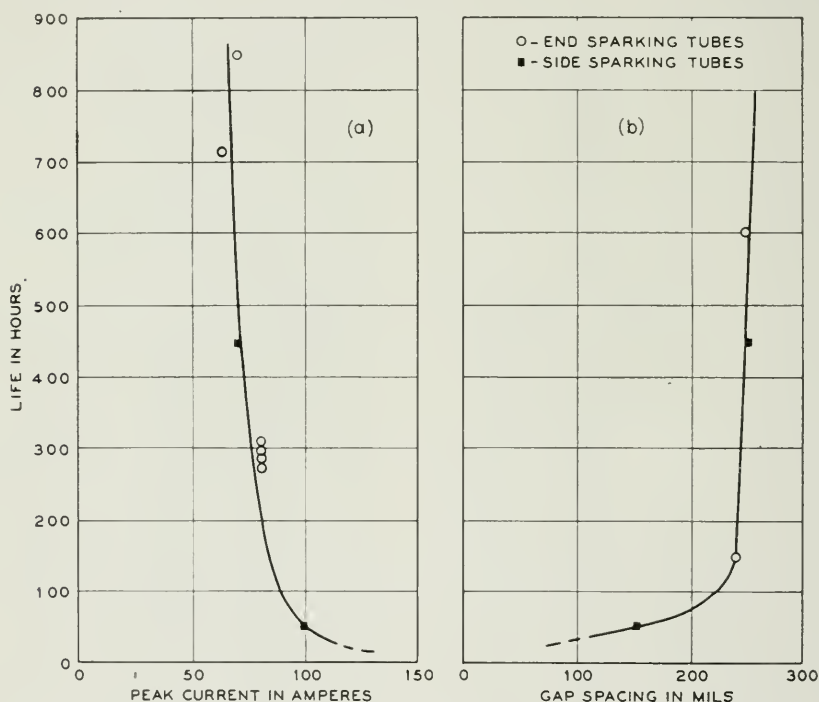
Fig. 13—Life in hours measured to the beginning of spike growth obtained with 5-micro-second pulses repeated 200 times a second (a) for a 60-mil gap and various peak currents, (b) for a fixed peak current of 70 amperes and various gap spacings.

we shall see in II-(e)—because of switching efficiency. This led to the development of a variety of unit gaps as described in II-(f).

(c) *The Mercury Cathode Gap*

Early in the study of the aluminum cathode gap it was realized that the sputtering difficulty might be largely if not entirely eliminated through the use of mercury as a cathode and the suppression of reverse current to avoid sputtering of the anode. It was shown that simple mercury pool cathode

gaps could switch peak powers in the megawatt range for long periods of time with stable operating characteristics. Under the urge for still higher powers than those which were handled by the aluminum cathode gaps, experiments were undertaken to develop a mercury cathode type of gap.

The main difficulty in the way of using mercury as a cathode is a mechanical one, as the conditions of operation of spark gap switches, particularly for

28" PRESSURE OF
75% HYDROGEN
25% ARGON

0.080" MOLYBDENUM

KOVAR CUP

0.310"
GAP

SINTERED IRON
POWDER SATURATED
WITH MERCURY
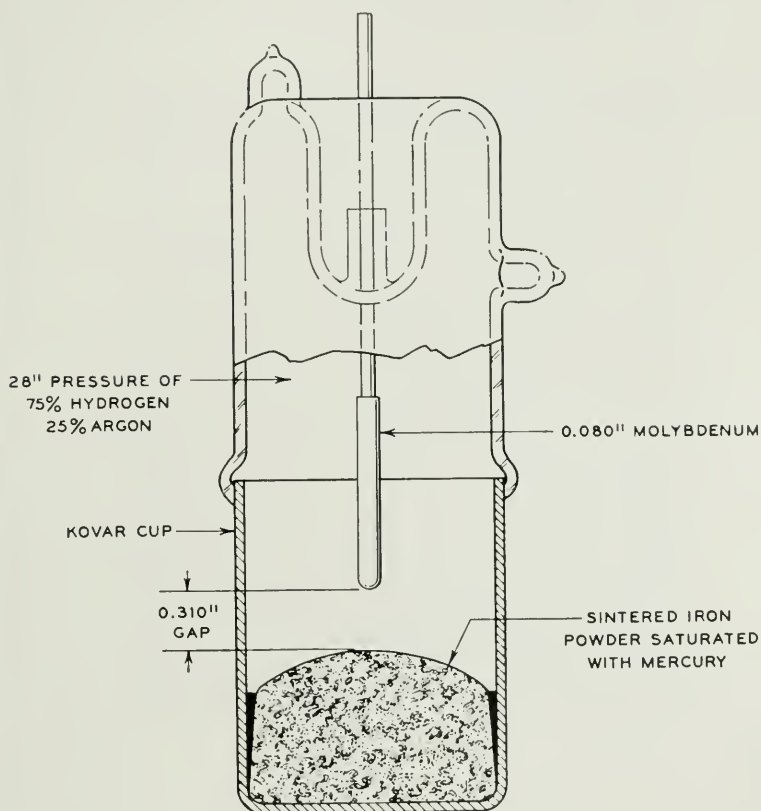
Fig. 14—Schematic drawing of an iron sponge mercury cathode unit gap—actual size.

airborne radar, demand that the sparking surface be rendered substantially quiescent. Preliminary experiments were made with metal baffles as damping agents and with metal wicks to furnish a mercury sparking surface. The latter led to the development of a sintered iron sponge saturated with mercury as the best means of obtaining a satisfactory cathode.

The constructional details of one of the earliest tubes of this type are given in the sketch, Fig. 14. The sintered iron sponge, a cross section of which is

shown in the photomicrograph Fig. 15, is about 60% porous. It was pre-
pared by pressing iron powder in the Kovar cup and sintering in a hydrogen
atmosphere. A special heat treatment to remove oxide made it possible to
fill all pores of the sponge with mercury and to supply a mercury film on its
surface. Under sparking conditions mercury from this film is evaporated
and is condensed on the tube walls, eventually returning to the cathode.
Due to capillary action the film is continuously replenished. This film
protects the iron sponge from sputtering provided that there is sufficient

Fig. 15—Photomicrograph (×15) of a section through a sintered iron sponge showing
porosity.

cooling of the cathode to maintain the mercury film at a temperature below
its boiling point.

These gaps are not temperature sensitive as are most electronic devices
containing mercury. This is because the mercury vapor plays no essential
role in the spark discharge, as indicated by the fact that dissipation measure-
ments—discussed in II-(e)—show its dependence on the hydrogen-argon
rather than on the nature of the cathode material. With adequate cathode
cooling gaps of this type operate satisfactorily over a range of ambient tem-
perature at least from −50°C to over 100°C. Practical gaps constructed

with iron-sponge mercury cathodes were developed to the manufacturing stage, as discussed in II-(f).

In addition to being capable of switching higher peak powers than aluminum cathode gaps, the mercury cathode gaps can be designed to have superior operating characteristics. Through the use of small radius anodes not possible with the aluminum cathode gaps, a wider operating range and much less time "jitter" can be attained. The small anodes build up corona at voltages less than those of breakdown, thus furnishing radiation prior to breakdown. For special applications, gaps have been developed having a range approaching 3 to 1 in a two-gap circuit, capable of switching 10 megawatts peak power, for many hundreds of hours, and having a time "jitter" of less than 0.02 microseconds at the operating voltage.[2, 3]

(d) *Starting and Operating Characteristics*

It has already been stated in II-(a) that starting and operating characteristics of series gaps cannot be interpreted simply because, under the circuit conditions of rapidly varying voltage, the breakdown voltage of a spark gap is not singly valued. Because of spark formation time the minimum voltage at which a spark gap will break down increases as the rate of rise of the voltage across it increases. Further, due to spark delay time, the voltage across the gap at breakdown is usually still higher than this minimum value. It is therefore impossible to designate a unique breakdown voltage of a spark gap when the voltage across it is increasing with time. It is, however, possible to find a practical minimum and maximum breakdown voltage for a particular rate of rise of voltage. The difference between this maximum and minimum value is a measure of the maximum spark delay time. It is for the purpose of reducing this spark delay time that corona points (or radium) are introduced, and it will be shown that the value of both spark delay time and spark formation time have an important bearing on the operational characteristics of fixed gaps.

In addition to rate of voltage rise, the breakdown voltage of a spark gap depends on the amount of ionization in the gap due to a previous spark. When a spark discharge stops, a column of highly ionized gas is left in the gap. Although this column is rapidly de-ionized by recombination and diffusion of ions, a lower breakdown voltage is found for many microseconds in consequence of this residual ionization. The minimum value of the breakdown voltage of the gap is therefore a function of the time

---

[2] F. S. Goucher, J. R. Haynes and E. J. Ryder, High Power Series Gaps Having Sintered Iron Sponge-Mercury Cathode, P.B. 19640, U. S. Department of Commerce, Office of the Publication Board.

[3] J. R. Dillinger, Operation of Sintered Iron Sponge-Mercury Cathode Type Series Gaps at S.C.I., A.E.W. and 5 Microsecond Conditions, P.B. 13270, U. S. Department of Commerce, Office of the Publication Board.

after the spark ceases and is called the re-ignition voltage of the gap. It will be shown that this re-ignition voltage determines to a large extent the starting voltage of the fixed gaps.

Before describing the sequence of events required for starting and operating, it is desirable to define our terms more precisely than we have defined them up to this point. The minimum operating voltage is the lowest switch voltage at which the tubes will continue to break down 100% of the time under the action of the trigger pulse, and the maximum operating voltage is that higher switch voltage at which spontaneous breakdown of the series of gaps never occurs. Thus the operating range of voltage is that which includes those voltages existing across the series of gaps, at the time of application of trigger pulse, for which the tubes always break down under the action of the trigger pulse but never before. Starting voltage is defined as the minimum value of d-c voltage at which a series of gaps can be made to break down under the action of the trigger pulse. Starting thus differs fundamentally from operating in that while operating demands that the series of gaps always breaks down under the application of the trigger pulse, starting requires only that the gaps break down once in many trigger pulses occurring in a fraction of a minute. Thus, a starting voltage is always lower than the minimum operating voltage. However, due to the doubling of the switch voltage when starting occurs, the d-c power supply voltage required to start may be higher than the d-c power supply voltage at the minimum.

The results of a quantitative oscillographic analysis of starting and operating characteristics of a pair of preproduction W. E. 1B22 tubes[4] are now presented in detail, for they are qualitatively representative of all spark gaps. These tubes operate in a two-gap circuit, a schematic of which is shown in Fig. 5 (a). The analysis is carried out by an examination of the voltage-time wave which occurs at the point of application of the trigger pulse, the midpoint of the two gaps. It will help in understanding the oscillograms[5] which follow if it is borne in mind that the voltage across gap 1 is the voltage shown on the oscillogram with respect to ground or "O" voltage, while the voltage across gap 2 is the voltage shown on the oscillogram with respect to the switch voltage.

The sequence required for starting is shown in Fig. 16 (a). Just before the application of the trigger pulse the voltage at the midpoint of the two gaps is half that of the applied d-c by virtue of the resistance divider. When the trigger pulse is applied, the voltage rises to A (3.8 kv) which is the minimum breakdown voltage for these tubes with voltage rates of rise encountered in the trigger pulse. Gap 1, therefore, may break down at A

---

[4] These tubes contain both corona points and radium to reduce spark delay time (see II-(e)).

[5] The time scales of these oscillograms are expanded in regions of very rapidly reversing voltage in order to make clear the sequence of events.
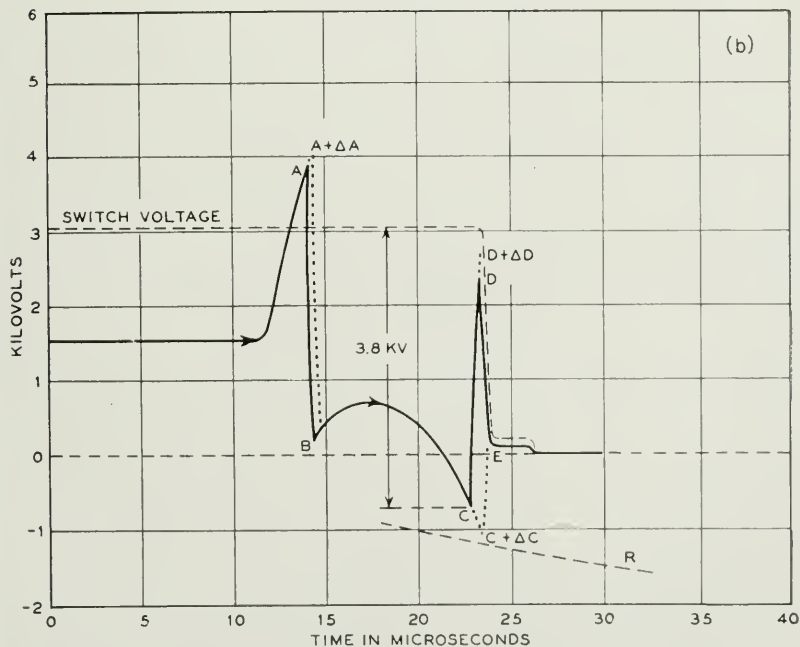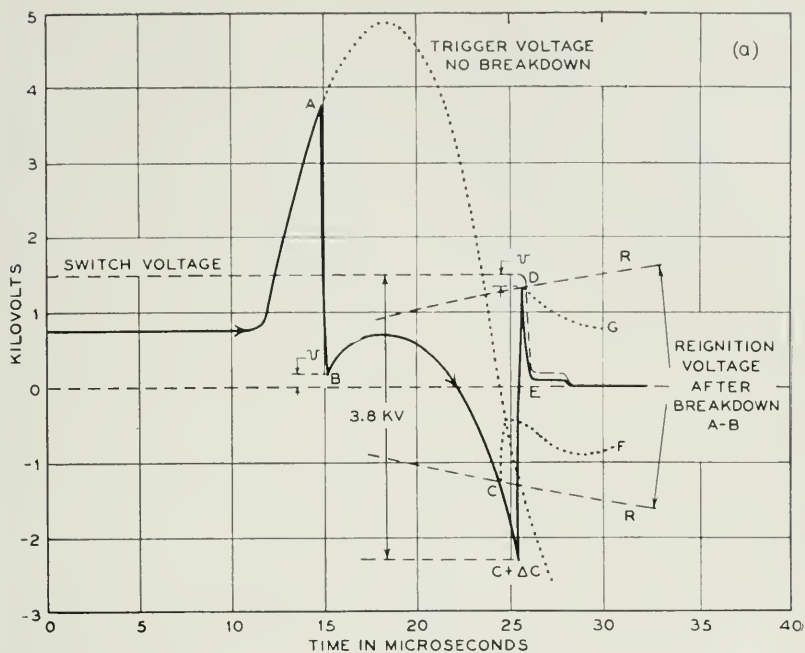
Fig. 16—Oscillographic traces of voltage vs. time as measured at the mid-point of a two-gap circuit during breakdown of 1B22 tubes, (a) for starting, (b) for operation at minimum switch voltage.

passing a low-energy spark supplied by the trigger circuit.   In consequence of this, the voltage drops sharply to B and then the discharge stops since the voltage (v) remaining is insufficient to maintain the discharge.   This voltage, called the extinguishing voltage, is about 0.2 kv for these low energy sparks.   Gap 1 is now ionized and has the independently measured re-ignition voltage characteristics, R, as shown.   Under the action of the trigger pulse the voltage then proceeds to $C + \Delta C$ when gap 2 may break down since it has the minimum required voltage across it (3.8 kv).   When this occurs, the voltage rises sharply to D, which falls short of the switch voltage by the amount of the extinguishing voltage (v).   At this point gap 1 may re-ignite. If this occurs both gaps are simultaneously conducting and the switch voltage drops to L while passing the high-current pulse of energy from the network.   This sequence occurs relatively infrequently.

Because of spark delay time, instead of breaking down at A, gap 1 may break down at some higher voltage, or not at all.   Instead of gap 2 breaking down at $C + \Delta C$, gap 1 may break down in the reverse direction at any voltage higher than C, its re-ignition voltage, and is only prevented from doing so by spark delay time.   Also, because of this delay time, gap 1 will usually fail to re-ignite at D, its re-ignition voltage, and since D is also the extinguishing voltage (v) for gap 2, the potential will drop to G under control of the trigger pulse.   If any one of these things occurs the gaps will not start on that particular application of trigger pulse.   However, since the pulses are applied at the rate of many hundred a second, it is usually only a fraction of a second until the desired sequence is obtained.

From the conditions essential for the consummation of each of the three steps necessary for starting, it follows that the starting switch voltage V_{dc} must be equal to $A - (R + \Delta C)$ or $v + R$, whichever is the greater.   Since R, the re-ignition voltage, increases with time, $A - (R + \Delta C)$ decreases while $v + R$ increases with time.   A minimum for V_{dc} will, therefore, be obtained when the period of the trigger voltage wave is such that when gap 2 breaks down,

$$A - (R + \Delta C) = v + R, \tag{1}$$

and since also for this minimum

$$V_{dc} = A - (R + \Delta C) \tag{2}$$

we get

$$V_{dc} = \frac{A - \Delta C + v}{2}. \tag{3}$$

By substituting the observed constant values of A, ΔC and v in (3) we get $V_{dc} = 1.5$ kv, which is the value of switch voltage depicted in the diagram. This diagram is, therefore, that for optimum period of the trigger voltage
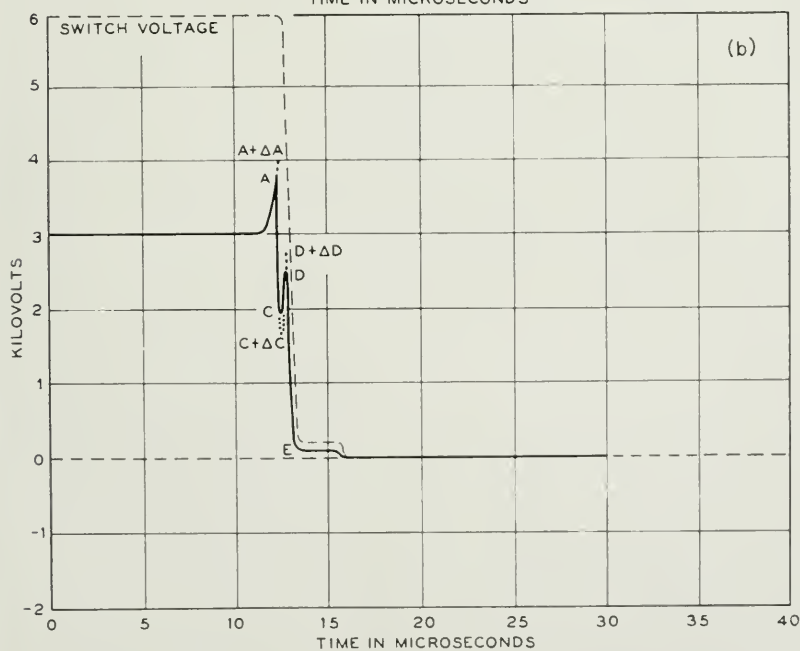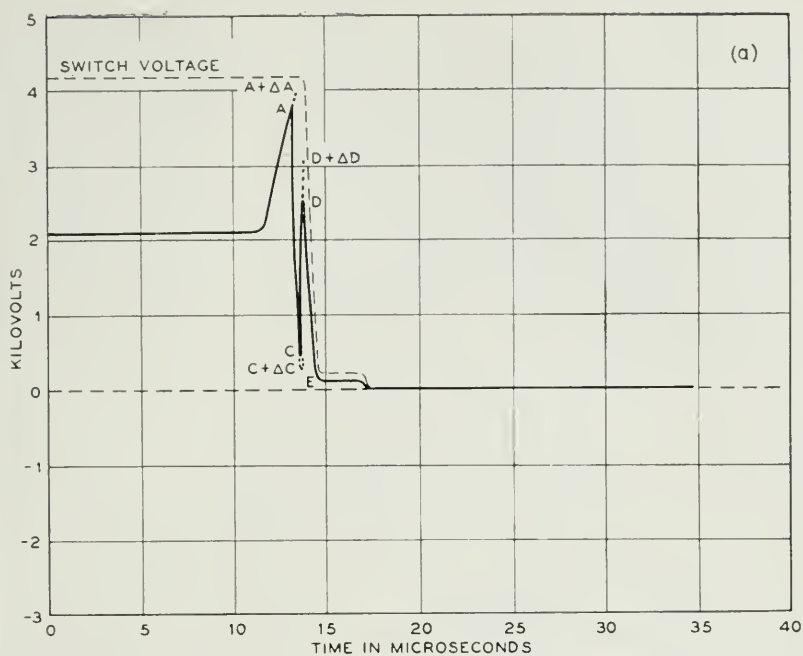
Fig. 17—Oscillographic traces of voltage vs. time as measured at the mid-point of a two-gap circuit during breakdown of 1B22 tubes (a) for normal operating switch voltage (b) for operation at maximum switch voltage.

wave.   That this is actually a minimum was demonstrated experimentally
by varying the period of the trigger pulse. V_{dc} increased for pulse periods
both greater than and less than that shown in the diagram.   The increase
was small and so is of no great practical interest, but it does confirm the pre-
diction made on the basis of the above analysis.

After the tubes have started the switch voltage is nearly double the d-c
voltage, and the tubes will operate continuously if the switch voltage is
above the minimum operating voltage.   The sequence of events near the
minimum operating voltage is shown in Fig. 16 (b).   During operation the
spark delay time is much less than during starting, as indicated by a smaller
voltage is established.
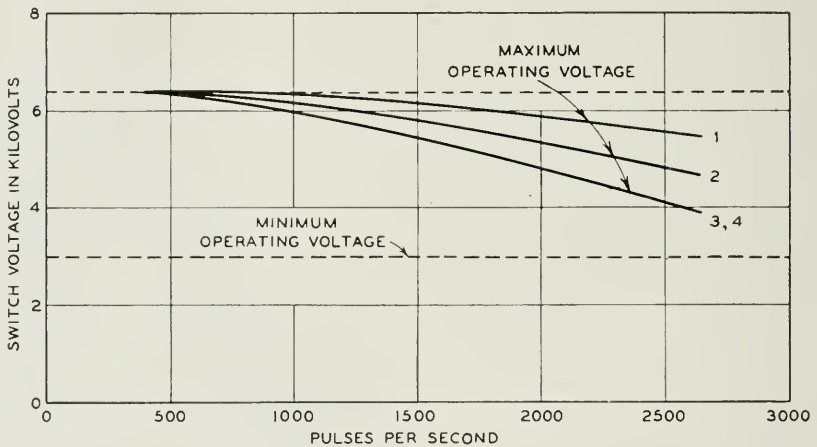
Fig. 18—Maximum operating voltage of 1B22 tubes in a two-gap circuit as affected by
pulse repetition rate for a variety of pulsing conditions as follows:

Curve	Pulse Duration in Microseconds	Load in Ohms
1	0.75	55
2	0.75	30
3	0.75	15
4	1.50	30

Of course, no spark gap tubes are designed to operate very close to their
minimum operating voltage.   A margin of safety is always maintained.
The characteristics of these tubes with a switch voltage at a practical operat-
ing voltage is shown in Fig. 17 (a).   Gap 1 breaks down between A and $A + \Delta A$
and before gap 1 is extinguished gap 2 breaks down between C and $C + \Delta C$.
Since in this case both gaps are conducting simultaneously, the main pulse
passes without re-ignition of gap 1.   The voltage at the midpoint of the
two discharges rises to a value between D and $D + \Delta D$, due to the rapid
change of spark impedance.   This sequence always takes place since ample
margin is provided.

If the switch voltage has been increased to a value near the maximum operating voltage, the voltage-time characteristic shown in Fig. 17 (b) results. Exactly the same sequence occurs as before. However, if the voltage be slightly increased above the value shown, the gaps can break down spontaneously during the network charging cycle and before the application of the trigger pulse, even though the value of A is some 20% greater than the charging voltage applied to the gap. This is the expected effect of spark formation time on minimum breakdown voltage since the rate of rise of trigger voltage is far higher than that of the network charging voltage. When spontaneous breakdown occurs, because of circuit conditions, both the rate of rise of the voltage of the network charging cycle and its peak value are increased. Since the switch voltage arrives at a higher value in a shorter time, spontaneous breakdown is most likely to occur again. The effect is cumulative so that, after a few increasingly frequent cycles, an arc is established. It is clear that this arcing must never be allowed to occur in the operating range.

These characteristics were taken while using a current pulse of 0.75 μs duration at a repetition rate of 1000 per second and a 30 ohm resistance load. This produced a peak current at the maximum operating voltage closely closely equal to the switch voltage divided by twice the resistance load, or about 100 amperes. Under these conditions, due to the relatively low pulse repetition rate, there is little residual ionization in the gaps at the time of the next pulse, so that the gaps have closely recovered their maximum breakdown voltage. However, as the pulse rate is increased, thus decreasing the time between pulses, the value of switch voltage at which the gaps break down spontaneously is found to decrease due to residual ionization. Thus the maximum operating voltage is a function of the pulse repetition rate.

The decrease of the maximum operating voltage as a function of pulse rate, for these tubes, is shown in Fig. 18 for a variety of pulsing conditions.

Curve 2 was obtained with the 0.75 μs pulse and a 30-ohm load. It will be observed that the maximum operating voltage decreases with pulse rate in the expected manner.

If the peak current of the pulse be decreased, fewer ions are produced in the spark and so at any given time after the pulse one would expect less residual ionization in the gaps. Curve 1 was obtained by keeping the pulse duration the same as before but increasing the load resistance to 55 ohms. Thus the current at a given switch voltage was reduced to 30/55 of its former value. It will be seen that, as predicted, the drop of maximum operating voltage with increased pulse repetition rate is less.

Conversely, if the current is increased the opposite effect is produced. Curve 3 was obtained by decreasing the load resistance to 15 ohms while keeping the pulse duration constant. This gives twice the peak current at

the same switch voltage as that of Curve 2 with a resultant increased residual ionization and a decrease of maximum operating voltage at the higher pulse repetition rates.

If, instead of doubling the current, the pulse duration be doubled, a similar increase in residual ionization is produced. Curve 4 was obtained by doubling the pulse duration (1.5 μs) and using a 30 ohm load. Thus, while the current is the same as Curve 2, the current pulse has twice the duration. It will be observed that for these pulses, doubling the time of pulse is the equivalent of doubling the current.

One might expect that the minimum operating voltage would also decrease as the pulse repetition rate is increased. However, experimentally it is found that, for these tubes, the minimum operating voltage is nearly constant and, therefore, independent of residual ionization. This result is produced largely because the maximum breakdown voltage of the gaps at the extremely high rate of voltage-time change encountered in triggering at the minimum is little affected by this amount of residual ionization.

Since the minimum is nearly constant the operating range of voltage of these tubes is a decreasing function of the pulse repetition rate, current, and pulse duration. This is in general true of all fixed spark gaps; however, the amount of decrease of operating range depends on the spark gap spacing, gas atmosphere and geometry of the electrodes.

(e) *Dissipation and Switching Efficiency*

In II-(d) we considered the voltage-time relationships leading to the simultaneous breakdown of series gaps. In this subsection we will consider the voltage and current relationships with time during this breakdown, and their bearing on spark dissipation and switching efficiency.

In Fig. 19 (a) are shown a voltage-time and current-time trace obtained oscillographically with a pair of 1B22 gaps. The voltage is measured across both gaps and corresponds to the dotted traces shown for switch voltage in Fig. 17 (a). The current pulse is shown in proper time relationship with the voltage trace. Similar traces are obtained for any pulse duration and peak current. These, then, may be considered as typical of all pulses produced by spark switching with these gaps.

In Fig. 19 (b) is plotted the impedance of both gaps with time, from which we see that the impedance of this switch falls rapidly in a small fraction of a microsecond to an average value of only a few ohms while the main current pulse is passing. The tail of the trace showing a negative impedance is due not to the gaps but to inductance inherent in their leads.

The solid trace, Fig. 19 (c), shows the product of voltage and current in kilowatts plotted against time. The integrated area of this plot corresponds to the dissipation per pulse of both gaps. This area is independent of the
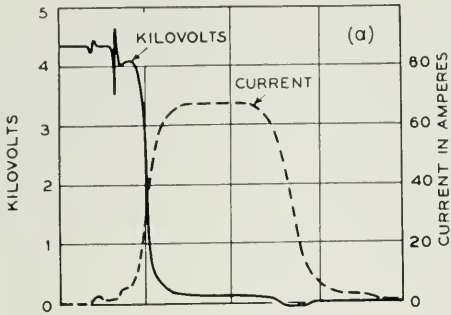
(a) Voltages vs. time and current vs. time for 0.75-microsecond pulse.
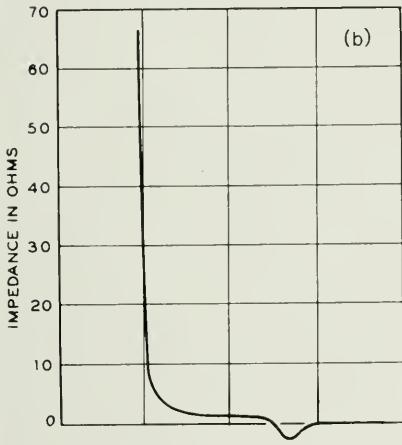
(b) Impedance vs. time.
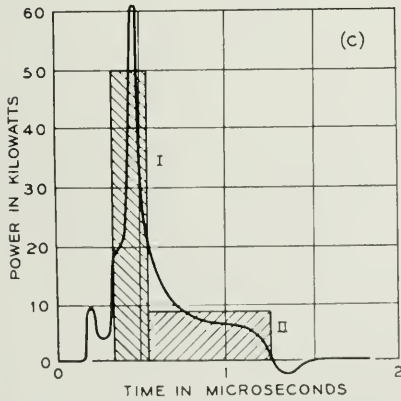
(c) Instantaneous power dissipated in gaps vs. time—solid trace from oscillographic, shaded areas from calorimetric measurements.

Fig. 19—Pulse characteristics of two 1B22 tubes operated in series.

pulse repetition rate, enabling one to determine the gap dissipation for any project by multiplying the loss per pulse by the repetition rate.

This area can be divided into two parts as suggested by the two shaded blocks I and II. The first part corresponds to the energy dissipated initially by the trigger and then by the pulse forming network in the brief transient period when the voltage across and the current through the gaps are changing rapidly. The former is comparatively small and usually can be neglected. The latter attains a maximum value of power when the impedance of the gaps approximates that of the load. The second so-called steady state part, corresponding to block II, represents the energy lost during the main pulse
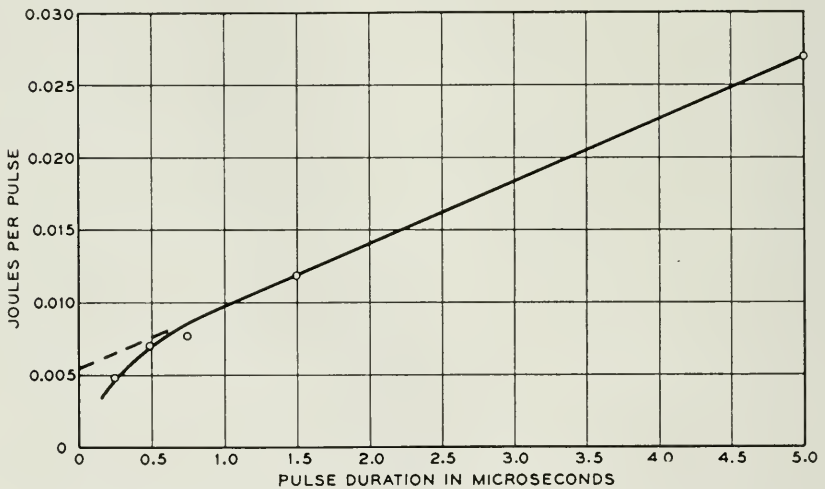
Fig. 20—Dissipation per gap per pulse vs. pulse duration for 1B22 gaps operated in series with a peak current of 70 amperes.

when the impedance of the gaps is low and comparatively constant. Its value will depend on both the pulsing conditions and the gaps themselves.

A calorimetric study was made of the dissipation of gaps as affected by various parameters. This method was superior to the oscillographic approach in that it afforded greater accuracy and ease of measurement. The curve, Fig. 20, shows observations in terms of joules per pulse per gap obtained calorimetrically with the 1B22 type tube as a function of pulse duration in microseconds. The peak current in all cases was 70 amperes and the trigger energy was included. It is clear that for pulse durations greater than 0.5 microseconds the dissipation D in joules per pulse per gap is given by

$$D = A + Bt \qquad (1)$$

where A is the intercept of the extrapolated solid straight line through the value of ΔA and ΔC. There are reasons for believing that this is due primarily to the higher electrode temperature, but is doubtless aided by residual ionization left over by the high-energy sparks now passing. As a consequence the first gap always breaks down at voltages intermediate between A and $A + \Delta A$. Gap 2 always breaks down at voltages between C and $C + \Delta C$, below the re-ignition voltage of gap 1, and gap 1 always re-ignites at voltages between D and $D + \Delta D$ allowing the main pulse of current to current to pass at a voltage E. However, if the switch voltage is decreased, $C + \Delta C$ will occasionally cross the re-ignition voltage characteristic R of gap 1. Gap 1 can then re-ignite and thus the gaps will not fire on the application of that trigger pulse. A second way in which the gaps can miss is by failure of gap 1 to re-ignite at D. Even though either of one of these events occurs only once in many thousands of pulses, a minimum operating observed points and where B is the slope of this line. The shaded blocks I and II of Fig. 19 (c) were obtained from values of the two terms A and Bt, respectively, showing graphically the agreement between the calorimetric and the oscillographic methods.

As a result of calorimetric measurements on a wide variety of gaps having either aluminum or mercury cathodes and operated under a wide variety of pulsing conditions, we have been able to establish an empirical formula for the dissipation D in joules per pulse per gap in terms of these gap parameters and pulsing conditions as follows:

$$D = 5.7(10)^{-7}I_pS + [40 + 3.9(10)^{-2}\, p^{\,0.4}S]\, I_pt. \qquad (2)$$

Here I_p is the peak current in amperes, S the gap spacing in mils, p the gas pressure of hydrogen-argon in inches of mercury, and t is the duration in seconds of an idealized square-top wave equivalent in ampere-seconds to the actual current wave. This formula holds for either aluminum or mercury cathodes and is independent of gap design. It is modified only slightly when pure hydrogen is substituted for the hydrogen-argon mixture, the constant $3.9(10)^{-2}$ becoming $3.1(10)^{-2}$. It is based on many measurements in which the parameters covered the following ranges:

Parameter	Range
S	40–350 mils
p	28–50″ Hg.
t	1–6 $\times$ 10^{-6} seconds
I_p	45–1070 amperes

After calculating the value of D from Equation (2) the dissipation in watts per gap for any project is obtained by multiplying by the pulse repetition rate. This equation does not include the trigger energy dissipated which usually

can be neglected but which can be measured independently and added if so desired.

It is to be noted that the first or transient term of the formula is unaffected by pulse duration and argon content and depends at least to a first approximation on only the peak current and length of spark. The numerical constant includes the time of this transient, the average gradient during this period, and a factor to reduce the peak current to an average value. The portion of the second or steady state term within the brackets represents the average voltage across a gap when it is highly conducting and is approaching the characteristics of a steady arc. This average voltage is separated into two parts. The first part, 40 volts, is the sum of the cathode and anode drops arising from space charges at the electrodes. The second part is the voltage drop along the positive column which has a pressure dependent uniform gradient and which is of the order of 100 volts per cm. It is only this gradient which is perceptibly altered when argon is added to the hydrogen.

From this formula it is possible to calculate the switching efficiency for any design of gap and set of pulsing conditions within the specified range of parameters covered by the formula. Calculation shows that with three gaps in series the switching efficiency in all projects was at least 90%, whereas with two gaps in series it was in most cases as high as 96%.

(f) *Development of Fixed Gaps for Manufacture*

The designs of the fixed gaps for manufacture were dictated by the requirements of particular modulators. Under the code number of each of the gaps a brief description is given of the electrical and mechanical requirements which had to be met.

W.E. 1B22

The 1B22 fixed gap tube is an aluminum cathode type with a hydrogen-argon filling. An exterior and a cross-sectional view are shown in Fig. 21. This fixed gap tube was developed for the modulator of an airborne radar known initially as ASH and later an AN/APS-4. In this modulator two tubes are used in series to switch a peak power of about 105 kilowatts into a W.E. 725A magnetron. It was desirable that the peak voltage in the modulator section be kept fairly low so that the circuit would perform satisfactorily at high altitude even when the pressurizing container was damaged. Furthermore, the equipment was to be very compact and light in weight.

In order to meet the requirements of this radar, two tubes were used in series with a peak switching voltage of 4 kilovolts. They were required to pass a current pulse of 67 amperes for 0.75 microseconds at two repetition

rates, one of 600, the other of 1000 pulses per second. They were also to operate for short periods at 2.25 microseconds and 330 pulses per second. The main problems in the design of such a tube were those of obtaining an adequate service life and a sufficiently low starting voltage.

As pointed out in II-(b), the life of an aluminum cathode gap of this type is critically dependent upon the anode-cathode spacing. For this tube a
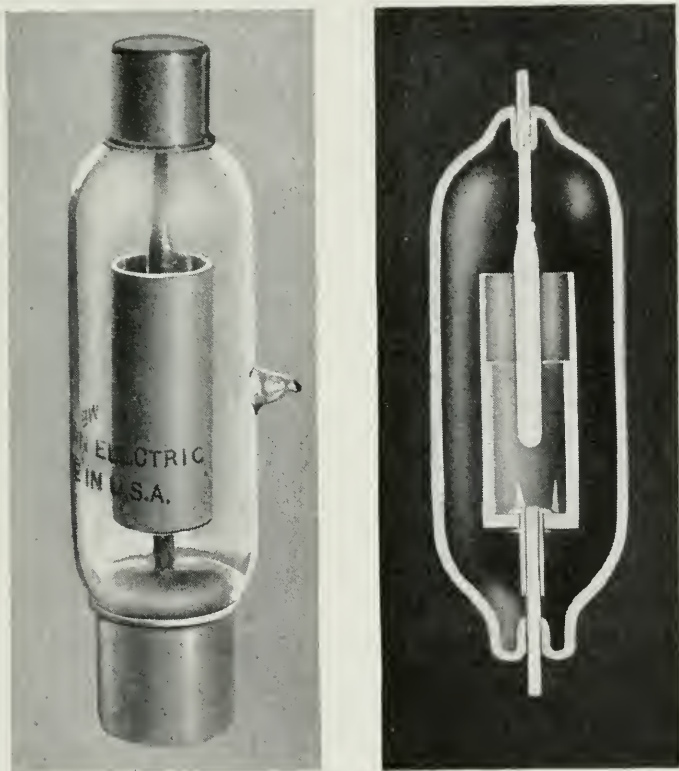
Fig. 21— Western Electric 1B22 spark gap tube.

spacing of approximately 150 mils was selected, the gas pressure being 20 inches of 75% hydrogen and 25% argon. This gave a life of about 500 hours for the 0.75 microsecond pulse, and a much shorter life for the 2.25 microsecond pulse. However, since the latter pulse duration is used only a small part of the time, the service life proved to be adequate. In order to obtain the maximum life from each tube, it was necessary that the anode and the cathode depart no more than a few mils from concentricity. Otherwise the sparking would not be uniformly distributed radially, leading to a

non-uniform anode build-up and a shortened life. Furthermore, in order to prevent failure of the tube, due to sputtered material destroying the insulation of the interior glass walls, the inside diameter of the cathode was enlarged near the open end, thus confining the sparking to the deeper portion of the cathode cylinder.

As discussed in II-(d) the starting voltage of a pair of fixed gap tubes is particularly important. The operating voltage of the tubes in this case is approximately 4 kilovolts which is derived from the resonant charging of the pulse shaping network condensers from a high voltage supply of about 2.2 kilovolts. The open circuit voltage of this supply is about 2.7 kilovolts. This, then, is the voltage available for starting the gaps. In order to make the gaps start at a voltage well below this value, corona points were introduced at the end of the cathode opposite the end of the anode, a small quantity of radium was also introduced in this region, and the anode diameter was reduced to the lowest value consistent with long life. The effectiveness of the corona points and the radium was reduced by the sputtered material during the life of the tube, but the irregular deposition of this sputtered material favored the production of corona and actually reduced the starting voltage to a lower value than that for a new tube.

The tube was designed for fuse clip mounting but it was found that the acceleration imparted to the tube when it was snapped into heavy clips was greater than that encountered in flying service. Accordingly, a special mounting was devised so that the tube would not be broken when being installed in the radar set. By the end of the war these tubes had been installed in approximately 15,000 radar equipments.

W.E. 1B29

The 1B29 fixed gap tube is similar in constructional details to the 1B22 except that it is smaller, the gap spacing being only 90 mils. An exterior and a cross-sectional view of the tube are shown in Fig. 22.

The gaps were designed to switch 2.8 kilovolts and to pass a peak current of about 27 amperes for 0.75 microseconds at a repetition rate of 2000 pulses per second. The main design problems were those of adequate life and stability of tube drop during conduction.

The small size of these gaps resulted in a life of only 300 hours which was, however, quite adequate for this application. As pointed out in II-(b) the argon was added to the hydrogen to ensure a uniform low impedance on sparking. The extremely small peak current required an increase in the amount of argon to 50% instead of the usual 25%.

In mechanical construction, the 1B29 is essentially a scaled-down 1B22. Because of the smaller size of the tube, no new problems existed in making it rugged.

Sufficient tubes were manufactured to supply approximately 3000 radar equipments.
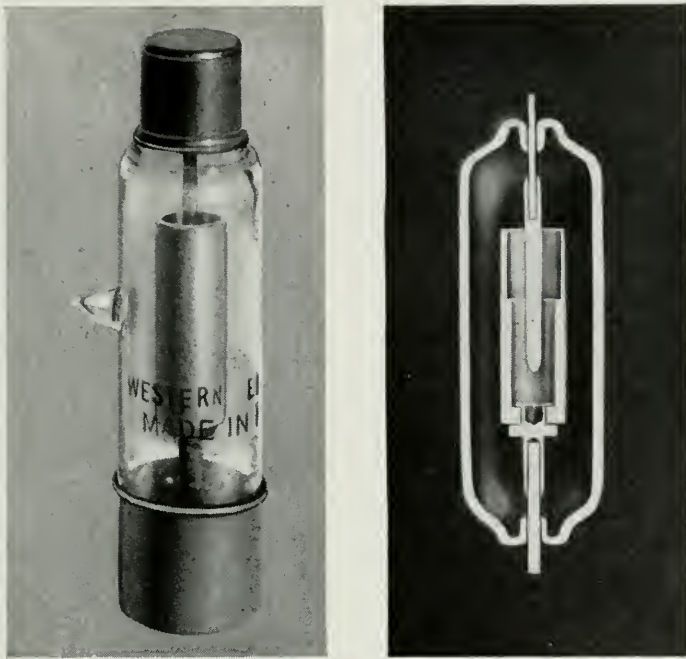
Fig. 22—Western Electric 1B29 spark gap tube.

W.E. 1B31

The 1B31 fixed gap was also an aluminum cathode gap, with a gap spacing of 300 mils and 24 inches of 75% hydrogen and 25% argon. An exterior and a cross-sectional view are shown in Figure 23. This gap was developed for an airborne radar. This modulator was required to furnish a peak power of 230 kilowatts to a W.E. 2J53 magnetron. The modulator was also to provide a range of pulse durations and repetition rates extending from 0.25 microseconds at 1600 pulses per second to 5.0 microseconds at 200 pulses per second.

In order to meet these requirements, two 1B31 tubes were used with a peak switch voltage of 8 kilovolts and a peak current of 75 amperes. By using a 300 mil spacing, a life greater than 500 hours was obtained at 200 pulses per second, 5 microseconds and 75 amperes. The other operating conditions were less severe from the life standpoint.

The wide spacing used meant a considerable increase in the size of the cath-

ode over the previous designs. To make this tube rugged, both electrodes were supported from large diameter kovar-to-glass seals. During assembly the cathode end of the tube was open so that a tool could be inserted to hold the anode concentric with respect to the cathode, while its supporting member was sealed to the glass. A cup was then brazed in to cover the cathode opening.
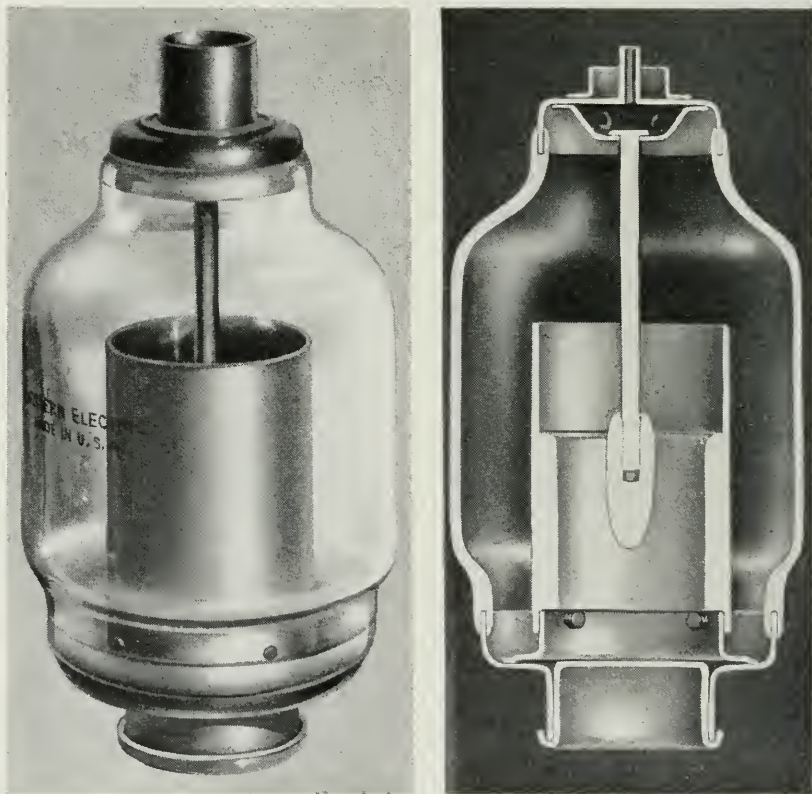
Fig. 23—Western Electric 1B31 spark gap tube.

Several hundred models of this tube were made in the laboratory and performed satisfactorily in the circuit. Due to circuit design changes, however, these tubes did not go into large scale manufacture.

W.E. 1B42

The 1B42 fixed gap tube departs considerably in design from the 1B22 and 1B29 in that mercury instead of aluminum was used as cathode. Its

construction is illustrated in Fig. 24.   This tube was developed for radars which were for long range search on shipboard.   In these modulators three tubes were used in series to switch a peak power of 0.8 megawatts and 1.4
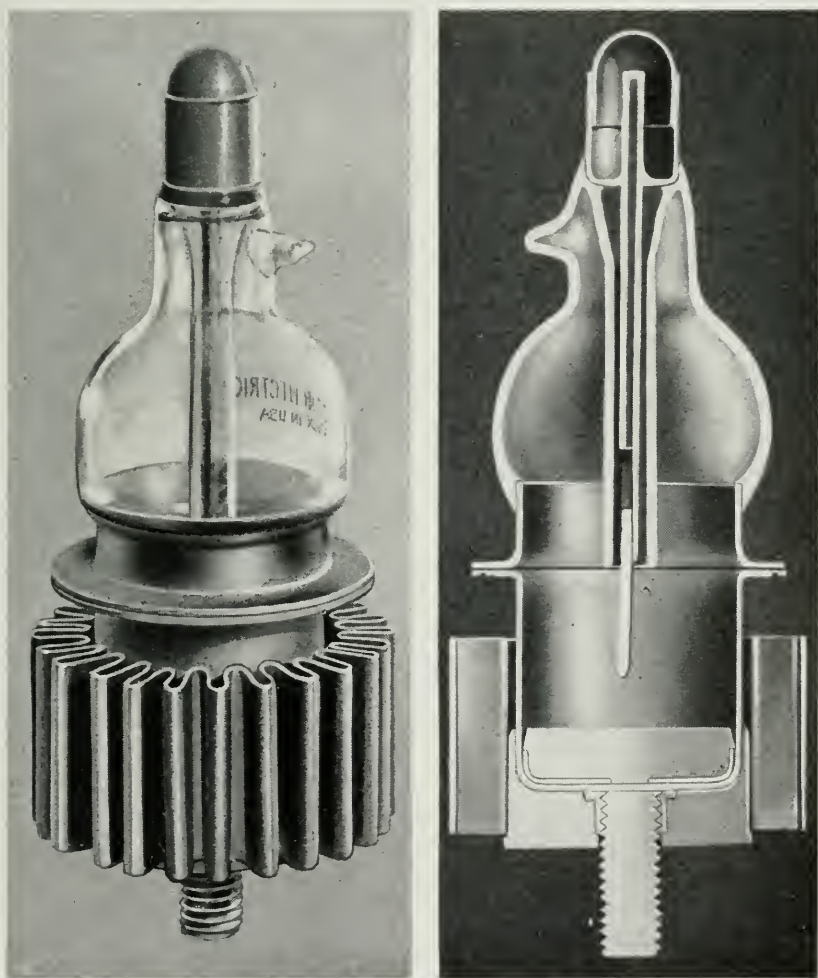
Fig. 24—Western Electric 1B42 spark gap tube.

megawatts, respectively, into high power triode oscillators.   These modulators were to be capable of operation at either half or full power.   The equipment was to be capable of withstanding the shock and vibration normally encountered on shipboard.

The series of gaps was required to operate with peak switch voltages

varying from 10.5 to 17.1 kilovolts, and to pass a maximum current of 200 amperes for 6 microseconds at a repetition rate of 180 pulses per second. They were also required to operate with 1.5 microsecond pulses at 600 pulses per second. The main electrical design problems were those of obtaining a wide voltage operating range and an adequate life with large peak currents and long pulses.

As discussed in II-(c), the use of an iron sponge mercury cathode with a molybdenum rod anode provided a wide voltage operating range as well as a long life with 200 ampere, 6 microsecond pulses. The mercury sponge cathode also met the vibration and shock requirements of shipboard operation.

In order to secure good wetting of the sintered sponge, which was essential to a long life, a special construction, as shown in Fig. 24, was used. The sponge was sintered directly into the bottom of a Kovar cup which had six radial vanes welded into it. These served to anchor the sintered material as well as to conduct the heat away from the center of the cathode. After the anode assembly and glass envelope were attached to the upper Kovar flange, the two sub-assemblies were welded together by means of a single ring weld. This allowed a minimum of handling of the sintered material and eliminated all glass work after the sintered material was inside the tube.

The processing of the tube consisted of first evacuating and then of heating the lower portion to 800°C while passing purified hydrogen through the tube. After the sponge had partially cooled, the mercury was introduced and wetting took place instantly.

Since the temperature of the center of the sponge must be kept below the boiling point of mercury, in addition to the internal vanes described, the Kovar cup was soldered into a block of copper to which was attached a folded copper radiator.

Several hundred models of the tube were made in the laboratory and delivered to the Navy and to equipment manufacturers. Full manufacturing information was turned over to the Navy which in turn issued a contract for the procurement of several thousand tubes.

Ratings

The ratings of the four different models of spark gap tubes developed by the Laboratories are summarized in Table 1. In order to permit the use of these gaps under a wide variety of operating conditions, yet prevent the simultaneous application of the maximum values of peak current, pulse duration, and repetition rate, a special system of rating was evolved. In addition to placing a maximum value on each of these three quantities a maximum value was also placed on the product of any two of these quan-

tities.   For instance, one of these products would prevent the use of very high peak currents along with very long pulses, a combination which would give a very short life, especially with aluminum cathode gaps.   Or another product would prevent the use of the tube at both high peak currents and high repetition rates, a condition which would not allow adequate deionization between pulses.   The later types of gaps were rated in this manner.

(g) *Evaluation of the Fixed Gap as a Modulator Switch*

In order to compare the performance of fixed gaps in radar modulators with that of other switching devices, as well as to assess their future possibilities, we may consider them with respect to the following points.

TABLE I

RATINGS OF W. E. FIXED SPARK GAP TUBES

Tube Type	Repetition Rate—pps		Peak Current a Max.	Pulse Duration µs Max.	Micro-Coulombs per Pulse Max.	Operating Voltage Range—2 gap Ckt. kv		Operating Voltage Range—3 gap Ckt. kv		Peak Trigger Voltage 3 gap Ckt. kv Nominal	Peak Trigger Voltage 3 gap Ckt. kv Nominal	Voltage Required for Starting kV	
	Min.	Max.				Min.	Max.	Min.	Max.			2 gap Ckt. Min.	3 gap Ckt. Min.
1B29	500	2100	30	0.75	—	2.6	3.0	—	—	3.0	—	1.9	—
1B22	300	1100	75	0.75	—	3.8	5.4	—	—	5.0	—	2.5	—
1B31	200	1600	300	5.0	375	7.3	9.2	—	—	8.0	—	6.1	—
1B42	160	1500	300	6.1	1280	9.0	11.4	10.5	17.1	10.0	15.0	6.5	8.5

(More complete information on the above tubes is contained in the JAN Specifications for individual tubes.)

1) *Peak current*—The present coded tubes cover a range of currents from 20 amperes to 300 amperes.   Experimental tubes have been tested up to 1000 amperes, and indications are that even larger currents are possible.

2) *Switch voltage*—The present tubes cover a range from 2.6 to 17.1 kilovolts.   Experimental tubes have been tested up to 30 kilovolts.

3) *Peak power output*—With the limits of peak currents and voltages on the present tubes, power outputs of 25 kilowatts to 2.2 megawatts are possible.   Experimental tubes were made which were capable of furnishing 15 megawatts.   Much larger power outputs seem possible.

4) *Pulse duration*—The maximum range of pulse durations covered by any of the present tubes is from 0.25 to 6 microseconds.   For pulses shorter than 0.25 microseconds the efficiency of these tubes would decrease rapidly.   Pulse durations much greater than 6 microseconds, however, could probably be used if proper attention is given to cooling.

5) *Pulse repetition rate*—A range of 160 to 2100 pulses per second has been covered by coded tubes. Experimental tubes with very short gaps have been tested up to 10,000 pulses per second. However, the design of tubes for practical operation in this region would entail considerable effort.

6) *Operating voltage range*—Although a given set of tubes may exhibit a wide range of operating voltage on a laboratory test, the rated range must be considerably less because of manufacturing variations and changes during life. However, since most radar modulators operate at a fixed power level, this limitation is not a serious one.

7) *Trigger requirements*—The spark gap tubes require a high-voltage low-current trigger supply. While this is more difficult to obtain than the low-voltage supplies required by some other modulators, it caused no real difficulty in practice.

8) *Time jitter*—Although the time jitter of coded tubes is of the order of one microsecond, experimental tubes have been made which have, at the operating voltage, a jitter of the order of one hundredth of a microsecond.

9) *Efficiency*—The switching efficiency for all of the past applications of fixed gaps has been in the range of 90 to 96 percent, which makes the fixed gap one of the most efficient switching devices for radar.

10) *Simplicity of manufacture*—Since the unit type of fixed gap has only two elements of simple geometry, its manufacture is relatively easy.

11) *Dependability*—The dependability of the fixed gap has been demonstrated by its satisfactory performance in its extensive application.

ACKNOWLEDGMENTS

Coil Pulsers for Radar

By E. PETERSON

R ADAR systems in current use radiate short bursts of energy developed by pulsing a high-frequency generator, usually a magnetron. One means of developing the requisite impulses employs a non-linear coil and is termed a coil pulser. Such pulsers are found in substantial numbers among the Navy's complement of precision radars. Most fire control radars on surface vessels are equipped with them, and all modern radar installations on submarines are so equipped for search and for torpedo control.

HISTORY OF DEVELOPMENT

Coil pulsers had their origin in the magnetic harmonic generators first built for the telephone plant. Multi-channel carrier telephone systems in general use throughout the Bell System require numbers of carriers, harmonically related in frequency. These are derived from non-linear coil circuits[1] which convert energy supplied by a sine wave input into regularly spaced, sharply peaked pulses.

When development was started on precision radars, one of these circuits generating a power peak of a few hundred watts, several microseconds in duration, was adapted to the purpose.[2] Its output was shaped and amplified by vacuum tubes of sufficient power to key or modulate the ultra-high-frequency generator of the radar transmitter. All early fire-control radars were made up in this way; hundreds are still in use.

The next development of pulsers for fire-control radars was directed toward higher-powered pulses, shorter in duration for good range resolution. These had to be provided by a small package pulser, small enough and rugged enough to mount integrally with the magnetron and the antenna. The power rectifier was to be located at any convenient distance, and the rectified voltage had to be low enough to permit the use of standard low-voltage cables. These requirements put high vacuum tubes at a disadvantage in handling the finally developed pulses. Pulse transformers had not attained their present state of perfection in dealing with short pulses at this early stage and the pulser therefore had to work the magnetron directly.

---

[1] Magnetic Generation of a Group of Harmonics, by Peterson, Manley and Wrathall, *B.S.T.J.*, vol. XVI, p. 437, 1937.
[2] Fire-Control Radars, by Tinus and Higgins, *B.S.T.J.*, January, 1946.

One arrangement developed by W. Shockley to meet these requirements used a thyratron as a switch to generate pulses. High vacuum tubes were used at low voltages for comparatively long-time intervals in the driving circuit. Deficiencies of the thyratrons available at that time prevented the generation of pulse powers as high as required. With the earlier experience on low-level coil pulsers in mind, it was natural to think of using a non-linear coil for switching pulses at high level, in place of the thyratron. Much development was required to arrive at suitable circuits embodying the basic ideas, to build non-linear coils capable of withstanding high voltages, to proportion the circuit elements for efficient operation at specified powers and pulse durations, and to shape the output pulse to the desired flat-topped form.

This development resulted in a power pulser mounted in an oil-filled steel box, with associated high vacuum tubes of the sturdiest sort mounted externally, operated from a 1200 volt d-c. supply. It was suitable for installation integral with the antenna, and rugged enough to withstand gun blast and shock. Life of the pulser box components is long, and performance stable with time and temperature. The time of pulse emission is linked precisely to the input wave, practically independent of voltage and frequency variations over a suitable range. Such precision timing, or freedom from jitter, permits starting the indicator equipment in advance of the pulse emission time so that target ranges may be accurately measured. The power rectifier voltage is much lower than that of the pulse applied to the magnetron, and the pulser works directly into the magnetron without requiring an intermediary pulse transformer.

Subsequent developments left unchanged the general form of the circuit and its mounting, but were devoted to achieving various pulse widths, powers, and pulsing rates to suit different applications. Pulse widths covered a range from two-tenths to over one microsecond, peak powers ranged from 100 to 1000 kw, and pulsing rates ranged from 400 to 3600 pulses per second.

Non-Linear Coil Structures

An idea of the general form and makeup of non-linear coils used in various radar developments can be had from the photograph of Fig. 1. All cores shown there are made of molybdenum permalloy tape, one mil thick. Insulation is electroplated on the tape in a silicic acid bath, and the tape is wound in ring form. After the standard magnetic anneal of 1000°C in hydrogen, the coating of insulation a fraction of a mil thick adheres firmly to the tape.

The smallest coil shown in Fig. 1 seen just in front of the oil filled container in which it is mounted is used for low-power pulse generation. Its core weighing 7 grams is wound on an isolantite form.

The two larger coils shown are used in power pulsers. Their cores are made up of self-supporting rings. The smaller coil has a core weight of one kilogram and is used at voltages up to 25 kv. for the generation of power peaks of the order of 100–250 kw. Phenol fibre is used to support and position the core and winding. The larger core has a weight of 13 kg. and is used at a voltage of 40 kv. in a pulser generating power peaks of one megawatt. Glass-bonded mica and built-up mica are used for support and
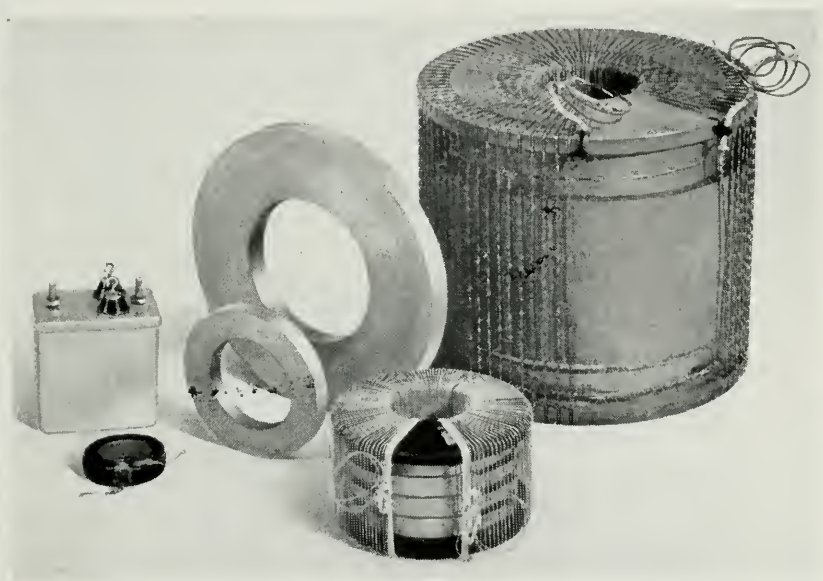
Fig. 1—Non-linear coils used in various radar transmitters. The smallest coil at the left, seen in front of its container, is used for low power pulse generation. The two larger coils are used in power pulsers developing 200 KW and 1000 KW peaks, respectively. The core rings of molybdenum permalloy tape are assembled into the coils shown.

positioning of the core rings and windings. The coils are assembled with other passive elements of the pulser network and the whole immersed in oil.

Operating principles of the two types of pulser circuits in which these coils are used are now to be discussed.

Low-Level Pulser

A schematic of the circuit used for developing low-power pulses is shown in Fig. 2a. Sinusoidal driving current (i_1) is introduced from the left, and a sharply peaked wave (i_2) is developed in the right-hand mesh. A resonant circuit (L_1C_1) serves to prevent dissipation of the generated pulse in the input mesh, and to tune out the input reactance at the driving frequency.

Capacitance and resistance elements (C_2R_2) in conjunction with the non-linear coil (L_2) make up the output mesh.

A complete cycle of the input wave is depicted in Fig. 2c, placed to correspond with the B-H loop of the non-linear core shown above it in Fig. 2b.
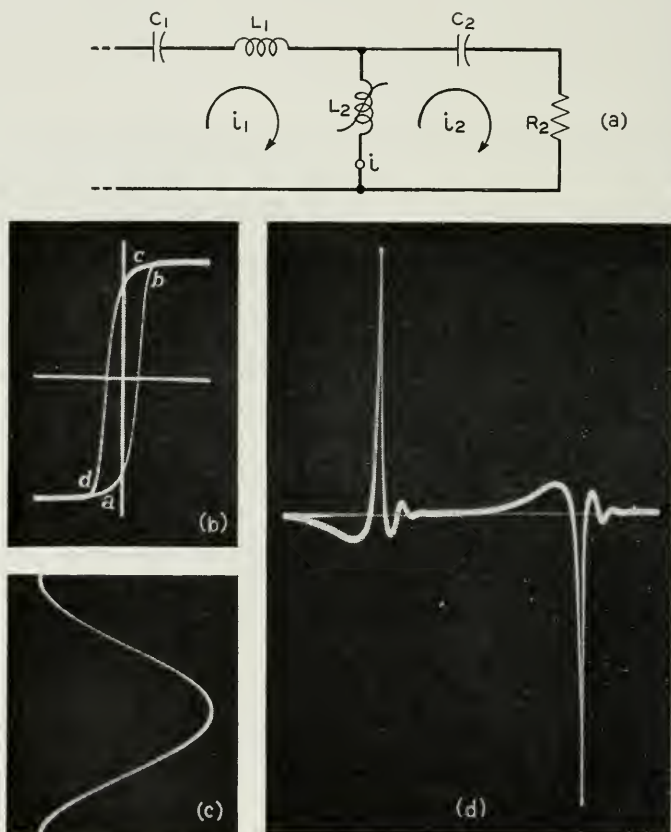
Fig. 2—Low-level coil pulser.
(a) Circuit diagram showing input tuning C_1L_1, non-linear coil L_2, output condenser C_2, and load resistor R_2.
(b) B-H loop of non-linear coil, with letters marking transitions between permeable and saturable regions.
(c) Sinusoidal input current wave scaled and placed to correspond with the horizontal scale of Fig. 2b.
(d) Pulsed output wave, i_2 as ordinate; i_1 as abscissa.

Action of the circuit is now to be followed throughout a cycle, starting with the input wave at its maximum negative excursion, condenser C_2 uncharged, and the core in its lower saturation region. Here the slope of the B-H loop and the corresponding differential permeability and inductance are

small. Hence the voltage drop across the coil is small. Little current flows in the output mesh, and practically all the input current flows through the coil. Matters are much different during the next interval in which the increase of current in L_2 brings the core into the permeable region a-b. Here the differential permeability is large so that part of the input current is diverted to the output mesh, charging the output condenser until upper saturation is reached at b. There the coil inductance falls to a low value, switching most of the condenser voltage across the load resistance. A current pulse accordingly develops in the output mesh lasting until the condenser charge is exhausted. The form of the current pulse shown in Fig. 2d approaches that of a highly damped sinusoid, and the pulse duration and magnitude are functions of the three elements of the discharge mesh. During the next half-period of the input wave, the same situation develops as in the first half-period, except that the corresponding currents and voltages throughout are reversed in sign.

According to this description the non-linear coil acts like a switch which automatically shifts the inductance from relatively high to relatively low values at specific coil currents. When the core is driven well into saturation, as is the case here, the ratio of these two inductances can be made large, usually in the neighborhood of several thousand. One feature of its action important from the efficiency standpoint is that the pulse occurs for the most part in the saturation region, where the contribution to eddy loss is small. The principal core loss occurs in the permeable region while the output condenser is charging, when variation of current through the coil occurs at a relatively slow rate.

In low-level radar applications the pulser output feeds a vacuum tube amplifier biased so that pulses of just one polarity are passed, the other oppositely poled pulse being cut off.

Since the range sweep of the radar receiver is initiated prior to pulse emission, the pulse should occur at a time linked precisely to the input wave. Otherwise the received pulse would be blurred introducing an uncertainty in measuring target range. No blurring (jitter) is visible with normal coil pulser operation. To get a measure of any variations which might be associated with core magnetization, tests were performed on a communication circuit in which jitter occurring at an audio rate would show up as noise. Measurements with a sensitive noise meter indicate the corresponding variation of pulse emission time to be smaller than 10^{-9} second.

Power Pulser

Operating Principles

The power pulser has the same type of discharge circuit as the low-level pulser just discussed. It differs in using a d-c. rather than an a-c. power
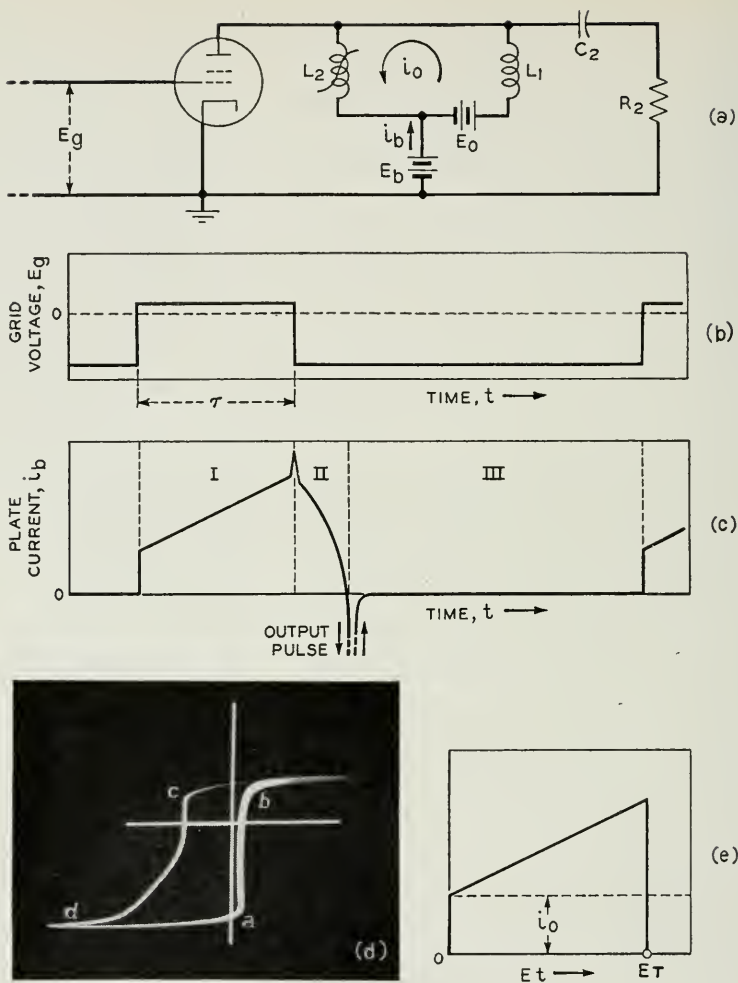
Fig. 3—Power pulser.

(a) Simplified circuit diagram showing charging tube at left, bias supply E_0, plate power supply E_b, linear coil L_1, non-linear coil L_2, output condenser C_2, and load resistor R_2.

(b) Rectangular wave of grid voltage impressed upon the tetrode of Fig. 3a. The tube conducts during the time τ in each cycle, and is cut off outside that interval.

(c) Plate current wave (i_b) corresponding to time scale of (b). During interval I, current is drawn through the paralleled inductors and the charging tube. At the end of this interval the tube is cut off and remains so until the start of the next cycle. During II, the magnetically stored energy is transferred to the condenser through R_2. At the same time the non-linear coil is brought toward saturation. During III saturation is reached; energy stored in C_2 is transferred to the load resistor through L_2 in a short pulse.

(d) B-H loop of non-linear coil used in the circuit of (a). Letters mark the most important transitions. During interval I magnetization proceeds from the lower left through a up to b; during II magnetization decreases past c down to d, and during III it extends far beyond the limits of the Figure to the left, returning to the neighborhood of d upon completion of the output pulse.

(e) Plot of current in linear coil during charging interval I against the product of coil voltage and time. Enclosed area represents energy stored in the linear coil. The rectangular area under the dashed line drawn through i_0 represents that part of the stored energy which varies with bias current.

608

source, and in charging the load condenser by a free, rather than by a forced oscillation. Energy for the free oscillation is taken from the d-c. source in a preliminary operation, in which energy is stored in a linear inductor. This preliminary operation consists in closing a d-c. path from the plate power supply through the linear inductor by means of a high vacuum tube, permitting current to build up with time. After a predetermined time has elapsed, the tube circuit is opened, the d-c. path is thereby interrupted, and energy stored in the inductor transfers to the load condenser. In this way the voltage to which the load condenser is charged can be made many times greater than the voltage of the plate power supply. The simplified circuit of Fig. 3(a) will serve to bring out salient operating features. Conduction of the tetrode at the left is controlled by a rectangular wave of grid voltage (Fig. 3b) developed by a multivibrator (not shown) which swings the grid from a potential below cutoff to one just above cathode potential. The plate power source E_b feeds two inductors in parallel, L_1 being linear, and L_2 non-linear. A small biasing voltage E_0 drives polarizing current i_0 through the two inductors in series.

The preliminary operation which serves to transfer energy from the main power source to the inductors is initiated when the tetrode grid is driven positive. Current from the main source builds up through the paralleled inductors and the tetrode as indicated on Fig. 3c, interval I. The region in which the non-linear coil works may be seen from the hysteresis loop of Fig. 3d. Its operating point is displaced to the left of the origin near d by the bias current. When the tetrode conducts, current in the non-linear coil rises rapidly at first in the lower saturation region until a is reached. The rise thereafter is comparatively small and slow in traversing the permeable region a-b, while at the same time current builds up in the linear coil at a much greater and practically uniform rate. When the core of L_2 reaches saturation near b its inductance again drops, preventing further rise of current in L_1. At this time the tetrode is driven below cutoff and remains out of the picture until the start of the next cycle.

The second interval, in which energy is transferred from the linear inductor to the load condenser, starts with the cutoff of tetrode current. This transfer is effected in an oscillation with frequency determined mainly by the paralleled inductors and the load condenser. In this interval II of Fig. 3c, current through the non-linear coil falls suddenly at first from b to c and then more slowly from c to d. The rate of change in region c-d is much greater than that in a-b as indicated by the fainter trace in Fig. 3d, so that eddy currents in the core are increased and the slope of the descending branch of the loop reduced correspondingly. Thus some of the energy previously stored in the linear inductor is used up in completing the magnetization cycle and this part, consequently, is not available for transfer to the load

condenser. The maximum voltage to which C_2 is charged in this interval is made much greater than that of the d-c. power source (E_b). The ratio of these two voltages depends upon the ratio of the inductance charging time in the preceding interval to the oscillation period. Both factors can be varied over wide limits, and step-up ratios of roughly ten to twenty are generally used.

The third interval starts with magnetization of the non-linear core near point d on the loop, where the inductance again drops. This situation is precisely the same as that previously described for the low-power pulser. As a result the condenser discharges through the load resistance at the time indicated in Fig. 3c, driving the core far into saturation with a field of the order of a hundred oersteds. This field extends too far to the left of point d to be shown in Fig. 3(d). Here the differential permeability approaches unity, and the correspondingly low inductance permits a rapid build-up of pulse current. Evidently but one pulse is produced each time the tetrode conducts, and the number of pulses produced per second is changed simply by varying the input frequency without requiring any circuit change, power dissipations permitting.

Energy storage in the linear coil depends upon its inductance, upon the bias current, and upon the peak current reached during the tetrode conduction interval. A plot of the current in L_1 against the product of time and of voltage across the coil permits this energy to be represented as an area (Fig. 3e). Evidently a given area can be made up by varying the relative sizes of its component triangle and rectangle, only the latter varying with bias current. If for example the bias is reduced to zero, the rectangle would vanish and the peak current would have to be increased to attain the original amount of stored energy. The higher maximum current requires more cathode emission of the tetrode and leads to greater plate power dissipation. Thus in addition to determining the energy stored, the amount of bias is one of the factors determining power dissipation capacity and emission which must be provided in the driving tube or tubes. Additional factors enter to make a bias corresponding to d (Fig. 3d) the most favorable from an efficiency standpoint.

The operating principles developed above in terms of a simplified circuit have been applied to a number of practical circuit forms which are described in the sections following.

Load Circuit

In radar applications the useful load is a magnetron which takes the place of the linear resistance previously considered. Since the magnetron viewed at its input terminals acts essentially like a negatively biased rectifier, additional means must provide for the flow of condenser charging current in a

direction opposite to that of the discharge pulse. This takes the form of a suitably poled diode shunted around the magnetron input terminals. After the main discharge pulse is completed, reactive elements are left with some little energy which tends to redistribute throughout the network. In course of redistribution, additional pulses of lower energy may occur shortly after the main pulse is completed. This tendency is a harmful one if the after-pulses are large, since echoes from short-range targets are obscured. Suppression of after-pulses is assisted by shunting around the diode-magnetron a linear inductance known as a clipping choke. This added inductance slows down the rate at which energy is redistributed, and permits the diode to fulfill its second function of dissipating the greater part of the residual energy. The shunting inductor, too, is made to fill a second function. Through provision of a bifilar winding, it passes heating current to the filament of the magnetron, thereby eliminating the need for high-voltage insulation otherwise required in the filament transformer.

Magnetic Bias

Several arrangements have been worked out for supplying various amounts of bias, some of them using a separate source, others being self-biased.[3] In general the use of external bias leads to a lower demand on the driving tetrode and is associated with pulse production at best efficiency. Circuits dispensing with an external bias source are that much more convenient in use, where the added tube demand and the lower efficiency corresponding can be handled without undue increase of the tube complement. In general the energy delivered to the magnetron is roughly 25 to 55 per cent of the plate energy input, with the higher figure applying to the higher outputs and external bias.

Transformer Coupling

In some cases it is convenient to equip the non-linear coil with primary and secondary windings providing voltage transformation or isolation to avoid adding a transformer for that purpose. The first case arises in the higher-powered pulsers, where the load condenser has to be charged to a voltage greater than the driving tetrode can withstand. For the Western Electric 5D21 tubes customarily used, voltage breakdown occurs near 20 kv, while condenser voltages in certain of the pulsers reach 30 and 40 kv. This situation calls for a step-up ratio from primary to secondary to fit the required potentials. The need for isolation may be illustrated by reference to Fig. 3a where the bias battery E_0 is shown maintained at the plate supply potential above ground. To avoid the resulting insulation problems in a

[3] One widely used circuit using a small amount of self-bias was developed by L. G. Kersta and E. E. Crump.

rectifier built to supply bias, a secondary winding is readily provided on the non-linear coil for connection to the linear coil and to the bias rectifier, which can then be maintained with one side at ground potential.

In either case whenever coupled windings are employed, the inside winding is invariably made to carry the discharge pulse. This provision results in minimum saturation inductance, since the inner winding is brought as close to the magnetic core as the voltage breakdown strength of the intervening dielectric permits. This winding is disposed as uniformly as possible around the core to avoid leakage which would add to the saturation inductance, and so limit the rate of current build-up in the pulse. The other winding can then be disposed with generous spacings, and with partial core coverage if desired.

Pulse Shaping

The oscillation frequency of the magnetron is determined primarily by its internal structure, although it is to some extent a function of the impressed potential. Departure of the driving wave from perfectly rectangular form permits the oscillation frequency to vary during the pulse, to an extent depending upon the size and duration of the departure and upon the characteristics of the magnetron.[4] Frequency modulation thus produced disperses energy over the spectrum. With the receiver band width limited to reduce noise and interference, one effect of this spreading of energy over the spectrum is to cut down the strength of the observed echo. For this reason, other things being equal, rectangularity must be approximated well enough to make the wasted energy a small fraction of that usefully employed.

It is convenient to regard the rectangular wave as synthesized by a series of harmonically related sine waves of appropriate magnitudes. The fundamental component according to this concept has a half period equal to the duration of the pulse, and the other components, progressively smaller in amplitude, have frequencies which are odd harmonics of the fundamental. In the low-power pulser with its rounded discharge wave the harmonic waves are quite small in amplitude. To approach the flat-topped discharge wave necessary in the power pulsers, harmonic components must be built up. This can be done by providing additional resonances in the discharge circuit at the wanted harmonic frequencies.

With the close spacing between circuit elements and their proximity to the pulser box walls, parasitic capacitances of appreciable magnitude add to those normally present. These involve dielectrics of low loss and, since the circuit elements and connecting wires are firmly fixed in position, they are fairly well reproduced. They can be used, therefore, in conjunction

---

[4] The Magnetron as a Generator of Centimeter Waves, by Fisk, Hagstrum, and Hartman, *B.S.T.J.*, April, 1946.

with added reactors of small size to provide harmonic resonances needed to shape the discharge pulse. These help to bring up the third and fifth, and in some cases higher harmonics.

Results after shaping are shown in Fig. 4 for two extremes of pulse width. The shorter pulse, roughly a quarter microsecond in average duration,
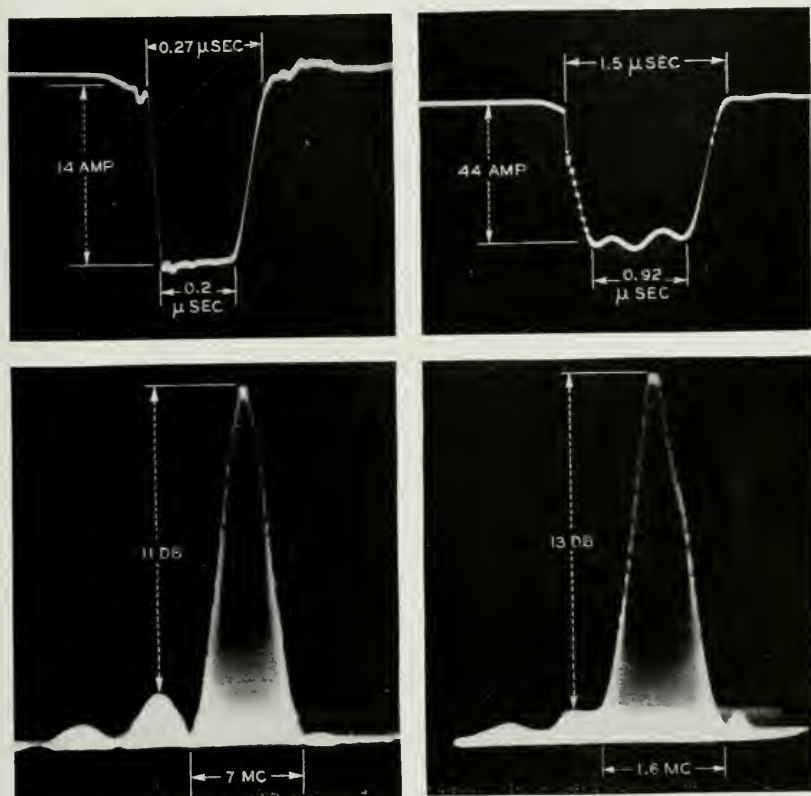
Fig. 4—Shaped magnetron current pulses, together with the radio frequency spectrograms corresponding. Pulse at upper left indicates presence of high harmonics; pulse at upper right shows strong fifth harmonic and little at higher harmonics. The band width of the main energy lobe, and the dispersion of energy outside that band in both cases indicate negligibly small effect attributable to frequency modulation.

evidences the presence of fairly high harmonics. The wider pulse, roughly one and a quarter microseconds in average duration, has a strong fifth harmonic and some even harmonics as well.[5] Below each pulse is shown a spectrogram of the corresponding magnetron high-frequency output, which

___

[5] Magnetron currents are shown rather than voltages, since current is a far more sensitive indicator of performance.

represents energy as a function of frequency. Different magnetrons are used with the two pulses; their operating frequencies and power capacities differ widely. Apparently frequency modulation exists in both cases to a small extent indicated by the departure of each spectrogram from symmetry about a vertical axis. Detailed study, however, shows that the band
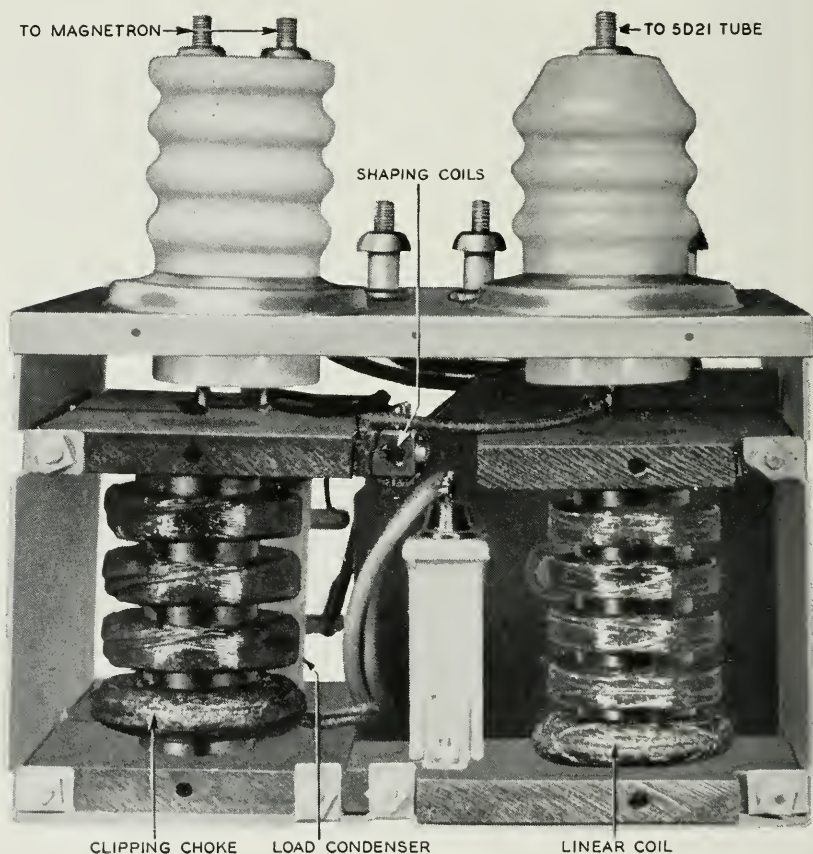
Fig. 5—Typical power pulser network.

width of the main energy lobe differs inappreciably from that with zero frequency modulation, and that the dispersion of energy attributable to frequency modulation is negligibly small. The pulses shown therefore provide satisfactory performance with their respective magnetrons.

Pulser Box

The form in which the typical power pulser network appears is shown in Fig. 5. Power peaks generated by the particular network shown are of the

order of 150 kw, with pulse durations of the order of a half microsecond. The non-linear coil here is similar to the one kilogram model pictured in Fig. 1; it is mounted on a panel back of the linear inductor indicated on the Figure. The two larger insulators are used to support high-voltage terminals, the double terminal at the left connecting to the cathode and heater of the magnetron and the single terminal at the right connecting to the tetrode plate. The smaller terminals provide lower-voltage connections including those to the plate power supply of 1000–1500 volts, the bias source where required, and the heating power supply for the magnetron. In use the network is sealed into a closely fitting oil-filled container.

Linear Servo Theory

By ROBERT E. GRAHAM

The servo system is a special type of feedback amplifier, usually including a mixture of electrical, mechanical, thermal, or hydraulic circuits. With suitable design, the behavior of these various circuits can be described in the universal language of linear systems. Further, if the servo system is treated in terms of circuit response to sinusoidally varying signals, it then becomes possible to draw upon the wealth of linear feedback amplifier design based on frequency analysis.

This paper discusses a typical analogy between electrical and mechanical systems and describes, in frequency response language, the behavior of such common servo components as motors, synchro circuits, potentiometers, and tachometers. The elementary concepts of frequency analysis are reviewed briefly, and the familiar Nyquist stability criterion is applied to a typical motor-drive servo system. The factors to be considered in choosing stability margins are listed—system variability, noise enhancement, and transient response. The basic gain-phase interrelations shown by Bode are summarized, and some of their design implications discussed. In addition to the classical methods, simple approximate methods for calculating dynamic response of servo systems are presented and illustrated.

Noise in the input signal is discussed as a compelling factor in the choice of servo loop characteristics. The need for tailoring the servo loop to match the input signal is pointed out, and a performance comparison given for two simple servos designed to track an airplane over a straight line course. The use of subsidiary or local feedback to linearize motor-drive systems, and predistortion of the input signal to reduce overall dynamic error are described.

1. Introduction

A SIMPLE servo system is one which controls an output quantity according to some required function of an input quantity. This control is of the "report back" type. That is, some property of the output is monitored and compared against the input quantity, producing a net input or "error" signal which is then amplified to form the output. The first statement defines the servo as a transmission system; the second, as a feedback loop. The problem of servo design is then to fashion the desired transmission properties while meeting the stability requirements of the feedback loop. This is the familiar design problem of the feedback amplifier.

2. The Servo Circuit

The design of linear feedback amplifiers has been developed to a high degree in terms of frequency response; that is, in terms of circuit response to sinusoidal signals.[1] The servo system is a special type of feedback amplifier, and usually can be made fairly linear. Thus, it is logical to analyze and design the servo circuit on a frequency response basis. Also,

[1] See "Network Analysis and Feedback Amplifier Design," by H. W. Bode, D. Van Nostrand Co., 1945.

616

servo systems usually are combinations of electrical, mechanical, thermal, or hydraulic circuits. In order to describe the behavior of these various circuits in homogeneous terms, it is desirable to recognize the analogous relationships established by similarity of the underlying differential equations. Before proceeding to a discussion of frequency analysis, a typical analogy between electrical and mechanical systems will be described.

2.1 *Electrical-Mechanical Analogy*

Confining the discussion to rotating mechanical systems, the analogy which will be chosen here puts voltage equivalent to torque, and current to rotational speed. This choice leads to the array of equivalents shown in Fig. 1; inductance, capacity, and resistance corresponding to inertia, compliance, and mechanical resistance, respectively. Charge is equivalent to angular displacement, and both kinetic and potential energy are self-analogous. The ratio of voltage to current, or torque to speed has the dimensions of resistance. In an interconnected electro-mechanical system, the ratio of voltage to speed or torque to current may be called a transfer resistance. Similarly, the ratio of voltage to angular displacement, or of torque to charge, is a transfer stiffness (reciprocal of capacity or compliance).

Some commonly used devices for coupling between electrical and mechanical circuits are shown in Figs. 2 and 3. The motor, Fig. 2a, is used to convert an electrical current i into a mechanical speed or "current" $\dot{\theta}$ ($= d\theta/dt$). The electrical control current i is produced by the voltage difference between an applied emf e and a counter-rotational emf (not indicated), acting upon the total electrical mesh resistance R_e.* In the mechanical circuit, a torque proportional to i forces a "current" $\dot{\theta}$ through the mechanical load R_m, J.

An equivalent mechanical mesh directly relating shaft speed to the applied emf is shown in Fig. 2b. A fictitious generated torque $\mu_t e$ acts upon the mechanical load through an apparent mechanical resistance R'_m. μ_t is a transfer constant determined both by the motor properties and the electrical mesh resistance R_e. R'_m is similarly governed and is inversely proportional to R_e.

The motor may be compared to a vacuum tube having an amplification factor μ_t and a plate resistance R'_m. However, the motor is usually much more a bilateral coupling element than the vacuum tube, due to the effect of the counter emf upon the electrical mesh.

The potentiometer, tachometer, and synchro circuit shown in Fig. 3 are all means for converting a mechanical quantity to an electrical one. All three are substantially unilateral coupling elements. The potentiometer

---

* R_e includes both the source resistance and the motor winding resistance.

delivers an output voltage e proportional to its shaft angle θ. Thus, the ratio of e to θ is a transfer stiffness constant S_t. The synchro circuit consists of a synchro generator connected to a control transformer, and delivers an output voltage e proportional to $\theta_1 - \theta_2$, the angular difference between
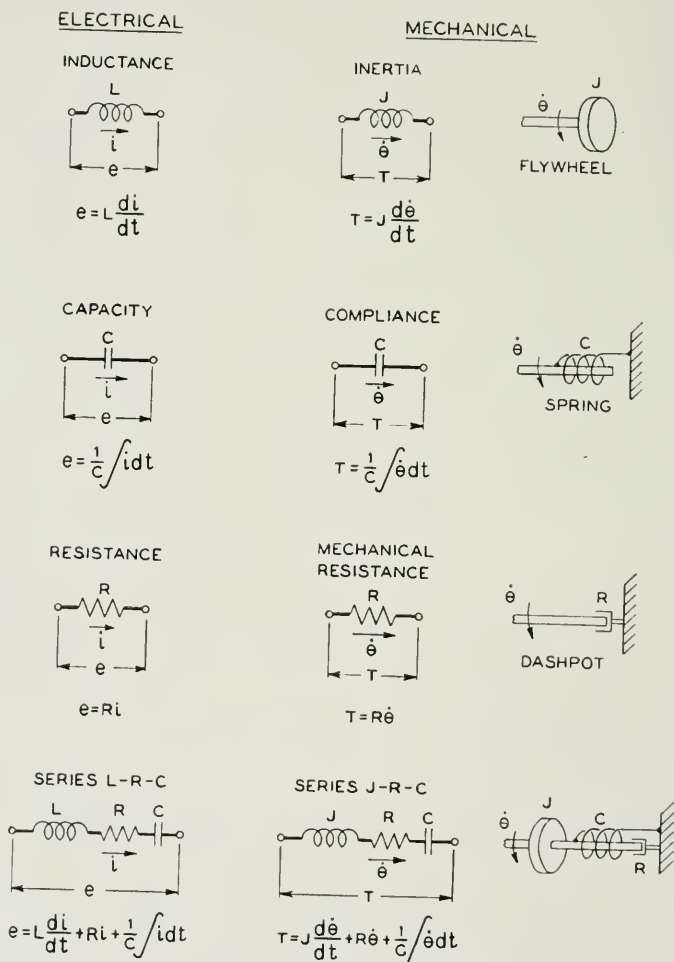
Fig. 1—Electrical-mechanical analogy.

the two shafts. Thus the action of the synchro pair is that of a combined transfer stiffness and differential. The tachometer is a generator which produces an output voltage e proportional to $\dot{\theta}$, the angular speed of its shaft. The ratio of e to θ is a transfer resistance constant R_t.

There are many other specific devices used to convert from mechanical to electrical quantities. Most are equivalent to the potentiometer or synchro circuit, one such being a lobing radar antenna, which delivers a voltage proportional to an angular difference. A different, less widely used device is the accelerometer, a generator which delivers an output voltage proportional to the angular acceleration of its shaft. Its characteristic is that of a transfer inductance or inertia.
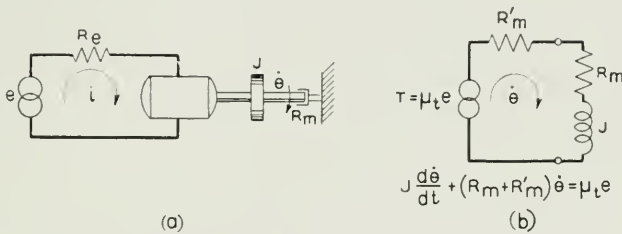
Fig. 2—Motor as a transfer device. (a) Motor and load. (b) Equivalent mechanical mesh.
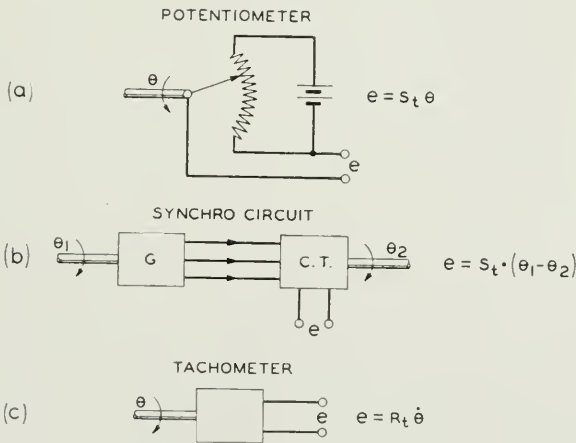
Fig. 3—Mechanical-electrical transfer devices.

2.2 *Frequency Analysis*

A brief review of the basic concepts of periodic analysis will be presented. It is assumed that the driving force applied to a network may be analyzed in terms of a series of sinusoidal components of various amplitudes, frequencies, and phases. The network response to each sinusoidal component is then evaluated, and the over-all result obtained by a summation of all such elementary responses. This is the formal procedure. Actually

it is often unnecessary to perform these precise operations in order to obtain a broad picture of the network behavior.

The method for determining the network response to a sinusoidal input is developed as follows. It is assumed that the circuit parameters are constant, independent of signal amplitude. Then, as indicated in Fig. 1, a single R-L-C or R-J-C mesh may be represented by a constant-coefficient linear differential equation. Choosing the electrical mesh for illustration,

$$(Lp + R + 1/Cp)i(t) = e(t),$$

where $p^n = d^n/dt^n$, $1/p = \int dt$, and $e(t)$, $i(t)$ are the mesh voltage and current respectively. This may be further abbreviated as

$$Z(p)i(t) = e(t), \tag{1}$$

where $Z(p) = Lp + R + 1/Cp$. For purposes of frequency analysis we are interested only in the forced or steady-state solution of (1), where $e(t)$ is a sinusoidal voltage $E \sin \omega t$. This steady-state solution is

$$i(t) = \frac{E}{|Z(j\omega)|} \sin (\omega t + \phi),$$

where $j\omega$ has replaced p in the function $Z(p)$, and the phase shift ϕ is the negative of the phase angle of the complex number $Z(j\omega)$.[2] This result is conventionally abbreviated as

$$I = \frac{E}{Z(j\omega)}, \tag{2}$$

where I is a complex number whose magnitude equals the peak amplitude of the current, and whose phase angle gives the associated phase shift. The function $Z(j\omega)$ is called the impedance of the mesh.

The relationship between torque, angular speed, and mechanical impedance is of course the same as expressed by (2). That is,

$$\dot{\theta} = \frac{T}{Z(j\omega)}, \tag{2.1}$$

where $\dot{\theta}$ is the complex peak amplitude of the sinusoidally varying speed, T is the peak amplitude of the applied sinusoidal torque, and $Z(j\omega)$ is the mechanical impedance obtained by substituting $j\omega$ for p in the operator $Z(p) = Jp + R + 1/Cp$. Since the angular displacement is the time integral

[2] ω is used to represent frequency in radians/sec, or 2π times frequency in cycles/sec.

of the speed, the expression for θ may be obtained by dividing both sides of equation (2.1) by p or, for the periodic case, by $j\omega$. Thus

$$\theta = \frac{T}{j\omega Z(j\omega)}. \tag{2.2}$$

The function $j\omega Z(j\omega)$ is the complex stiffness of the mechanical mesh. The phase shift of θ relative to $\dot{\theta}$ is -90 degrees, as seen from a comparison of (2.1) and (2.2).

For an electro-mechanical network consisting of a number of interconnected meshes, a set of simultaneous differential equations of the type of (1) may be written. If $j\omega$ is substituted for p in these equations, there results a set of simultaneous algebraic equations which lead directly to the steady-state periodic solution. If a sinusoidal voltage or torque is applied at some mesh of the network, the resulting sinusoidal current or speed response in some other mesh is given by, using the notation of (2),

$$(\text{Response}) = \frac{(\text{Cause})}{Z_t(j\omega)},$$

where $Z_t(j\omega)$ for the chosen pair of meshes is obtained from the solution of the algebraic equations. $Z_t(j\omega)$ is called a transfer impedance, and may express the ratio of a voltage to current or speed, or of a torque to current or speed. The above relation also may be written as

$$(\text{Response}) = Y_t(j\omega) \cdot (\text{Cause}),$$

where $Y_t(j\omega) = 1/Z_t(j\omega)$ is called the transfer admittance between the two chosen parts of the network. In this form the response amplitude is obtained by multiplying the forcing sinusoid by $|Y_t(j\omega)|$, while the phase shift is given directly by the angle of $Y_t(j\omega)$. The transfer ratio between like or analogous quantities at two parts of the network is similarly a complex function of frequency, having the dimensions of a pure numeric.

Servo systems usually consist largely of elementary networks isolated by unilateral coupling devices (vacuum tubes, potentiometers, etc.). Thus, over-all transfer ratios often may be evaluated by taking the product of a number of simple "stage" transfer ratios, rather than by solving a large array of simultaneous equations. If the absolute magnitudes of the transfer ratios or "transmissions" are expressed in decibels[3] of logarithmic gain, both the over-all gain and phase shift of a number of tandem stages may be obtained by simple addition of the individual stage gain and phase values.

The transfer ratios of the conversion devices shown in Figs. 2 and 3

[3] The gain in decibels for a given transfer ratio is taken to be 20 $\log_{10}$ of the absolute value of the ratio.

may be written by inspection.   Referring to Fig. 2b, the transfer admittance of a motor with resistance and inertia load may be written as

$$\frac{\theta}{E} = \frac{\mu_t}{j\omega J + R_m + R'_m},$$

$$= \frac{\mu_t}{J} \cdot \frac{1}{j\omega + \omega_m}, \tag{3}$$

where

$$\omega_m = \frac{R_m + R'_m}{J}.$$

ω_m is the reciprocal of the time-constant of the motor and control mesh, and is 2π times the "corner" frequency at which the inertial impedance just equals the apparent mechanical resistance.   Writing the transfer characteristic in terms of shaft position, rather than speed,

$$\frac{\theta}{E} = \frac{\mu_t}{J} \cdot \frac{1}{j\omega(j\omega + \omega_m)}. \tag{3.1}$$

For values of ω small compared with ω_m, θ/E is proportional to $1/j\omega$. This factor has a phase shift of -90 degrees and approaches infinity as ω approaches zero.   This is merely a statement in frequency analysis language that the motor shaft angle is the time integral of the applied voltage, for slowly changing voltage.   For more rapidly varying voltage, such that ω is large compared to ω_m, θ/E is proportional to $1/(j\omega)^2$ or $-1/\omega^2$, the angular variation being shifted -180 degrees with respect to the voltage variation.

The transfer ratio of the potentiometer, Fig. 3a, may be written as

$$\frac{E}{\theta} = S_t, \tag{4.1}$$

while for the tachometer, Fig. 3c,

$$\frac{E}{\dot{\theta}} = R_t, \tag{4.2}$$

or

$$\frac{E}{\theta} = j\omega R_t. \tag{4.3}$$

Usually at some point in the system a compensating or "equalizing" network will be included to modify the transfer ratio of the basic components to the desired over-all transmission characteristic.   Frequently this equalizer is incorporated in the electrical section of the servo because

of the comparative ease with which electrical circuit components may be assembled in desired combinations. The transfer characteristic of the equalizer may be simple or complicated, but in general may be written in the form,

$$\mu_e \sim \frac{(j\omega + \omega_1)(j\omega + \omega_3) \cdots}{(j\omega + \omega_2)(j\omega + \omega_4) \cdots}, \tag{5}$$

where the constants ω_1, ω_2, etc. may be real or complex. The synthesis of equalizing networks is a well known art and will not be discussed here, particularly since most of the equalization characteristics used in present servo systems can be realized with simple networks.

2.3 *Simple Servo System (Single Loop)*

The simple servo system may be divided into two basic parts, an amplifying circuit and a monitoring or comparison circuit. Such a division is

LOOP TRANSMISSION $= \mu\beta$

TRANSFER CHARACTERISTIC:

$$F_2 = \frac{\mu}{1 - \mu\beta} F_1$$
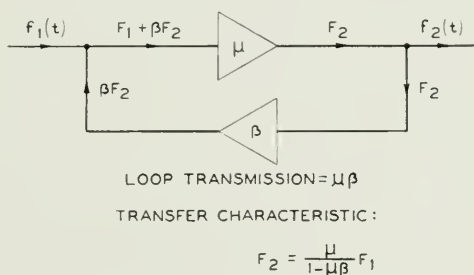
Fig. 4—Simple servo system.

indicated in Fig. 4, where μ and β are the transfer characteristics of the amplifying and monitoring parts, respectively. F_1 and F_2 represent typical sinusoidal components of the total input and output quantities $f_1(t)$ and $f_2(t)$,[4] while μ and β are complex-valued functions of $j\omega$ as described in the previous section.

The return signal βF_2 from the monitoring circuit is added to the servo input F_1 to form a net μ circuit input $F_1 + \beta F_2$. The servo transfer characteristic is found by setting

$$F_2 = \mu(F_1 + \beta F_2),$$

from which

$$F_2 = \frac{\mu}{1 - \mu\beta} F_1. \tag{6}$$

[4] That is, F_1 and F_2 are complex quantities employed in the same fashion as I in equation (2).

The closed system formed by the two basic circuits in tandem is of course a feedback loop, the loop transmission characteristic being given by $\mu\beta$.

Any desired form of servo transfer ratio may be obtained by an unlimited number of μ and β circuit combinations.[5] However, the β characteristic, which is usually determined by a passive network or an inherent property of a monitoring device, tends to be more stable with time and varying signal amplitude than that of the μ circuit, which may include vacuum tubes, motors, and other variable components. Consequently, it is desirable to have the servo transfer characteristic largely dependent upon the β circuit properties alone. This may be accomplished by making the loop trans-
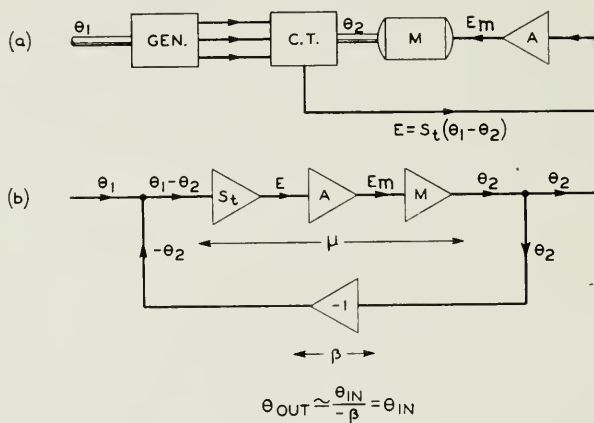
Fig. 5—Synchro follow-up system.

mission $\mu\beta$ very large compared to unity over the essential frequency range of the servo input signal $f_1(t)$. Under this condition, equation (6) becomes,

$$F_2 \simeq \frac{F_1}{-\beta}, \qquad |\mu\beta| \gg 1. \tag{6.1}$$

Thus the external transfer characteristic is set by β.* If, for instance, F_1 and F_2 are similar or analogous quantities and it is desired to have the servo output a replica of the input, β may be chosen as -1, yielding $F_2 \simeq F_1$.

It is not always easy to determine the basic parts μ and β of a servo by inspection of a schematic diagram of the system. An example is furnished by the synchro follow-up system shown in Fig. 5a. As previously discussed, the characteristic of the synchro comparison circuit is that of a differential

---

[5] Feedback stability requirements place certain restrictions on the permissible forms of $\mu\beta$. This will be discussed in the next section.

* The error arising from the approximate nature of (6.1) will be discussed in the next section as one type of "servo error."

transfer stiffness S_t, the voltage output of the control transformer being given by $S_t(\theta_1 - \theta_2)$. However, as seen from the modified diagram of Fig. 5b, the β characteristic is simply -1, the transfer constant S_t appearing in the μ circuit. Thus if the loop transmission is kept large, the essential relation demanded between θ_1 and θ_2 does not depend upon the value of S_t

$$\theta \simeq \frac{E}{-\beta} = \frac{E}{S_t} = C_t E$$
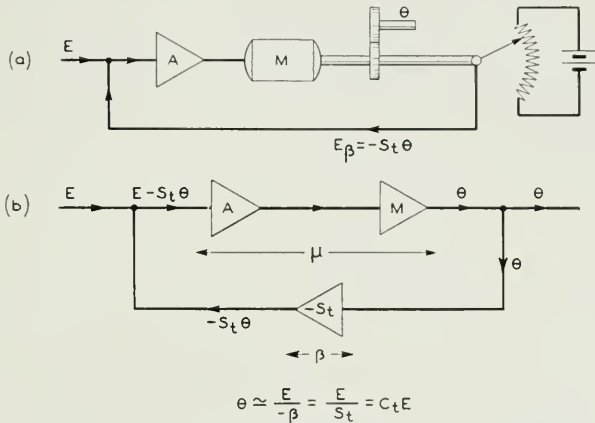
Fig. 6—Potentiometer loop.

$$\dot{\theta} \simeq \frac{E}{-\beta} = \frac{E}{R_t}$$
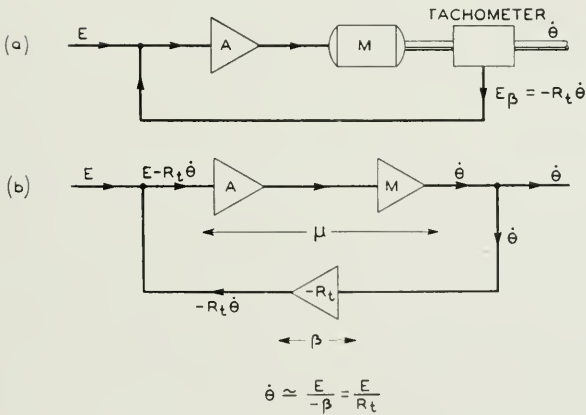
Fig. 7—Tachometer loop.

(as is obvious from physical considerations). This result also applies to a radar angle-tracking loop, where the received deviation or error signal is proportional to the difference between the angular coordinate of the target and that of the antenna system.

Figures 6 and 7 represent two servo systems where the input is electrical and the output mechanical. In Fig. 6 a potentiometer is used as a monitor-

ing device, the transfer stiffness in this case appearing in the β circuit. If θ is regarded as the output, then $\beta = -S_t$,* and the transfer characteristic is, for high loop gain,

$$\theta \simeq \frac{E}{-\beta} = \frac{E}{S_t} = C_t E,$$

where $C_t = 1/S_t$. Thus the over-all characteristic between input voltage and angular displacement is simply a transfer compliance constant. In Fig. 7 a tachometer monitor is used. Regarding angular speed $\dot{\theta}$ as the output, then $\beta = -R_t$, the transfer resistance of the tachometer. The transfer equation is thus

$$\dot{\theta} \simeq \frac{E}{-\beta} = \frac{E}{R_t} = g_t E,$$

where $g_t(= 1/R_t)$ is a transfer conductance.

3. Design of Simple Linear Servo Systems

The majority of servo systems in use, while often greatly extended in space and frequently including highly diversified transmission elements, may be represented by one essential feedback loop. However, a well designed servo often will incorporate numerous subsidiary or local feedback loops around stages of the system, in order to obtain a desired degree of linearity or performance with easily obtainable circuit components. Common examples of such local feedback loops are electrical feedback around vacuum tube amplifiers, and tachometer (velocity) feedback around motor-drive systems. These subsidiary feedback loops are almost always designed so that they are individually stable when the over-all feedback loop is opened (assuming the method employed for opening the over-all loop does not disturb the impedance terminations of the local feedback stages). If it is thus assumed that any subsidiary loops are individually stable, then the primary servo loop design may be treated simply as that of a single loop, whose over-all loop transmission is obtained by taking the product of the external transfer ratios of the various stages.

The design of a single loop servo may be divided into the design of the loop transmission $\mu\beta$, and one of the remaining parts, μ or β. As previously described, it is usually desirable to fix β according to the required basic input-output relationship of the servo, thus leaving $\mu\beta$ as a single characteristic to be chosen.

---

* It is assumed here that the transmission of the μ circuit is basically positive. The negative signs associated with S_t of Fig. 6 and R_t of Fig. 7 are then introduced (by poling) to make the loop transmission $\mu\beta$ essentially negative. This stipulation ensures what is commonly called "negative feedback," when the loop delay is zero.

As usual, specification of the form of $\mu\beta(j\omega)$ is beset by a series of performance objectives on the one hand and a set of physical limitations and restrictions on the other. Assuming the relationship expressed by (6.1) to be the required one, it would seem desirable to make $\mu\beta(j\omega)$ very large compared to unity at all frequencies. However, there are reasons why this is neither possible nor actually desirable. As the value of ω is increased, the loop transmission is eventually controlled by parasitic circuit elements such as distributed capacity and inductance in the electrical circuits, and distributed inertia, compliance, and backlash in the mechanical circuits. The effect of these parasitic elements at the higher signal frequencies is to cause $|\mu\beta|$ to decrease as a very high order of $1/\omega$ with increasing frequency. It will be shown, however, that feedback stability considerations require the loop transmission to be decreasing comparatively slowly through the frequency region where $|\mu\beta|$ is of the order of unity. Thus $\mu\beta$ must be reduced below unity at a frequency sufficiently low to avoid excessive contribution from the parasitics.

The presence of "noise" or undesired disturbances in the servo input signal is another compelling factor in the design of the loop characteristic. Input noise is harmful both in causing spurious output fluctuations and in overloading the power stages of the servo system. Both of these effects are reduced by narrowing the frequency band of the servo transfer characteristic. Referring to the expression for the transfer characteristic given by (6), it may be seen that a restricted transfer bandwidth may be obtained by reducing μ and $\mu\beta$ well below unity at a small value of signal frequency.[6]

On the other side of the picture is the requirement of fidelity in maintaining the desired input-output relationship. Undue narrowing of the transfer bandwidth of the servo results in large dynamic error, the magnitude of which depends both upon the character of the input signal and upon the chosen transfer characteristic.

The optimum design of a servo system, for a specified input signal and noise, thus is a compromise between dynamic error and output noise fluctuations, with stability considerations and parasitic circuit elements restricting the possible choice of loop transmission characteristics.

3.1 *Stability of Single Loop Systems*

The word *stable* as applied to a servo system is used here to imply a system whose transient response decreases with increasing time. It is possible

---

[6] When the β characteristic is under suitable design control, another method is available. Thus if β is made to rise in the frequency region of the desired transfer cut-off, and if $\mu\beta$ is maintained large beyond this region, (6.1) shows that the desired restriction is effected. For a given transfer characteristic, this cut-off method requires a wider frequency range for $\mu\beta$ and is thus more vulnerable to the effects of parasitic circuit elements. However, shaping of both the μ and β circuits permits a more rapid cut-off of the servo transfer characteristic than is possible with μ circuit shaping alone.

to determine the stability of a completed servo design by obtaining the transient solution of its differential equation. Though often very tedious, this is fairly straightforward. However, this method of procedure often is of little help either in guiding the initial design or in predicting the necessary changes, should the trial design be found unstable. The addition of even one circuit element to a design will generally create an entirely new differential equation whose solution must be found.

An alternative method for determining servo system stability, based on frequency analysis, furnishes the necessary information in a form which greatly facilitates the design procedure. This method is relatively simple to apply, even when the system has a large number of meshes and a high order differential equation, and the additive effects of minor circuit modifications are easily evaluated.

The stability of a single loop servo system—or of a primary loop, when the subsidiary loops are individually stable—may be investigated by plotting the negative of the loop transmission, $-\mu\beta(j\omega)$, on a complex plane for real values of ω ranging from minus infinity to plus infinity. (The negative sign is introduced because the loop transmission $\mu\beta$ is generally arranged to have an implicit negative sign, apart from network phase shifts. Thus $-\mu\beta$ is a positive real number when the network phase shift is zero.) *Then the necessary and sufficient criterion for system stability is that the resulting closed curve must not encircle or intersect* the point $-1,0$.* This type of plot is commonly called a Nyquist diagram, and is widely used in the design of electrical feedback amplifiers. An added stipulation is necessary if $\mu\beta(j\omega)$ becomes infinite at a real value of ω, say ω'. In this case an infinitesimal positive real quantity ϵ must be added to $j\omega$; that is, the function to be plotted is $\mu\beta(j\omega + \epsilon)$. This has no effect upon the plot except in the neighborhood of the singularity, where $\mu\beta(j\omega + \epsilon)$ is caused to traverse an arc of infinite radius as ω is varied through the value ω'.

As seen from (3.1), inclusion of a motor in a servo loop of the type shown in Figs. 5 and 6 will cause an infinite loop transmission at $\omega = 0$, assuming there is transmission around the remainder of the loop at zero frequency. The motor is the only commonly encountered circuit element capable of producing an infinite loop transmission at real frequencies.

In order to illustrate the use of the Nyquist diagram, a motor servo system of the type shown in Fig. 6 will be chosen. Again referring to equation (3.1), it may be seen that the transfer ratio of the motor and potentiometer is,

---

* This criterion is due to H. Nyquist, "Regeneration Theory,". *B. S. T. J.*, January 1932. Detailed descriptions of stability criteria for single and multiple loop systems are given by Bode, loc. cit., and by L. A. MacColl, "Fundamental Theory of Servomechanisms," D. Van Nostrand Co., 1945.

$$\frac{E_\beta}{E_m} = \frac{-S_t \mu_t}{J} \cdot \frac{1}{j\omega(j\omega + \omega_m)}.$$

It is assumed that the amplifier includes an equalizing network such that the over-all amplifier characteristic is

$$A\left(\frac{j\omega + \omega_1}{\omega_1}\right)\left(\frac{\omega_2}{j\omega + \omega_2}\right)^3,$$

where A, ω_1, and ω_2 are positive real constants. The loop transmission is given by the product of these two transfer factors and thus may be written as

$$\mu\beta(j\omega) = -\frac{\omega_0}{j\omega}\left(\frac{\omega_m}{j\omega + \omega_m}\right)\left(\frac{j\omega + \omega_1}{\omega_1}\right)\left(\frac{\omega_2}{j\omega + \omega_2}\right)^3, \tag{7}$$

where ω_0 is a positive real constant given by

$$\omega_0 = \frac{AS_t \mu_t}{\omega_m J} = \frac{AS_t \mu_t}{R_m + R'_m}.$$

The three factors in parenthesis have been so grouped that they all approach unity for small values of ω. Thus the low-frequency behavior of $-\mu\beta$ is given by $\omega_0/j\omega$. This quantity has a pole at $\omega = 0$, so that it is necessary to plot the function

$$-\mu\beta(j\omega + \epsilon) \simeq \frac{\omega_0}{j\omega + \epsilon}, \tag{7.1}$$

in the neighborhood of $\omega = 0$. As ω is, in succession, a small negative quantity, zero, and a small positive quantity; the expression of (7.1) is correspondingly a large positive imaginary, a large positive real, and a large negative imaginary. Thus (7.1) traverses an infinite arc from the positive imaginary axis to the negative imaginary axis as ω increases through the value zero.

Assuming the numerical values

$$\omega_0 = 200 \text{ sec}^{-1}$$
$$\omega_m = \quad 1 \quad \text{``}$$
$$\omega_1 = \quad 10 \quad \text{``}$$
$$\omega_2 = 200 \quad \text{``} ,$$

equation (7) may be rewritten as

$$-\mu\beta(j\omega) = \frac{200}{j\omega}\left(\frac{1}{j\omega + 1}\right)\left(\frac{j\omega + 10}{10}\right)\left(\frac{200}{j\omega + 200}\right)^3. \tag{7.2}$$

The phase angle of $-\mu\beta$ in degrees is, from (7.2),

$$B = -90 - \tan^{-1}\omega + \tan^{-1}\frac{\omega}{10} - 3\tan^{-1}\frac{\omega}{200}, \tag{7.3}$$

while the absolute magnitude is given by

$$| \mu\beta | = \frac{200}{\omega} \sqrt{\left(\frac{1}{1 + \omega^2}\right)\left(\frac{10^2 + \omega^2}{10^2}\right)\left(\frac{200^2}{200^2 + \omega^2}\right)^3}. \qquad (7.4)$$

The Nyquist diagram of (7.2) is shown in Fig. 8. To emphasize the important features, radial magnitudes have been plotted on a logarithmic
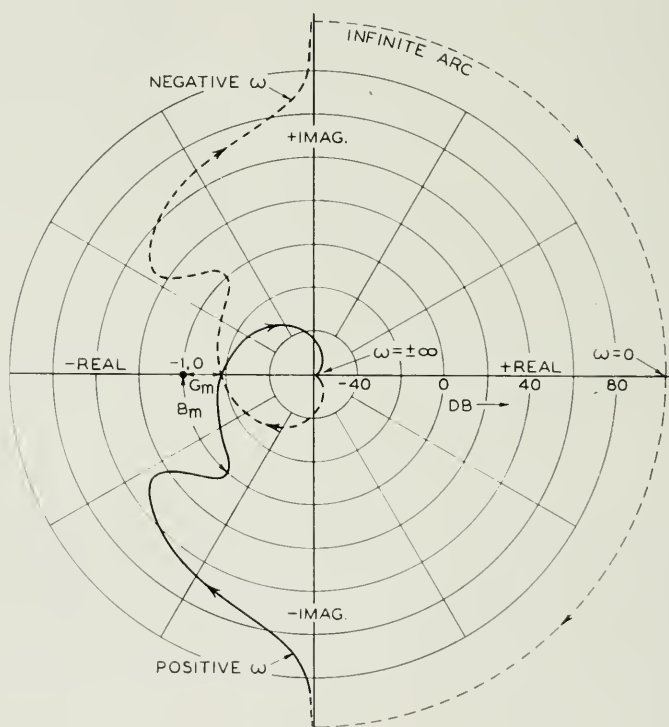

Fig. 8—Nyquist diagram of $-\mu\beta$.

scale.[7] The arrows indicate the direction of traversal as ω is varied from $-\infty$ to $+\infty$. The infinite arc traversed as ω varies through zero is indicated symbolically by the dotted semicircle in the right half plane.[8] As is the case for any physical system, the plot for negative values of ω is simply the mirror image of the positive frequency plot.

Since the polar plot does not encircle or intersect the "critical" point $-1,0$,

---

[7] Except in the immediate neighborhood of the origin, where a linear scale must be employed to plot the value $\mu\beta = 0$.

[8] The exact shape of this arc is of no consequence.

the system is seen to be stable.[9]   From a practical standpoint it is necessary
to know not only that a design is stable, but that it has sufficient margin
against instability.   The need for proper stability margin arises from two
general considerations.   First, the loop transmission of the physical system
will vary with time due to aging, temperature changes, line voltage fluctua-
tions, etc.   Also the physical embodiment will depart from the paper de-
sign due to errors of adjustment and measurement, and to the effects of
unallowed-for parasitic elements.   Second, a design which is too near
instability will have an undesirable transient response—large overshoots
and persistent oscillations—and will unduly enhance noise in the input
signal.

Stability margin is measured in a sense by the minimum displacement
between the polar plot and the point $-1,0$.   In feedback amplifier design,
two numbers often are taken as a measure of margin against instability.
These are called the *gain margin* and the *phase margin*.   The gain margin,
G_m, measures the amount, in decibels, by which the magnitude of $\mu\beta$ falls
short of unity, at a phase angle of ±180 degrees.   The numerical value of
gain margin for the system of Fig. 8 is about 18 db, which is the required
increase in amplifier gain to make the servo unstable.[10]   That is, this in-
crease in amplifier gain would multiply the curve of Fig. 8 by a constant
factor such that it would intersect the point $-1,0$.   The phase margin,
B_m, is equal to the absolute magnitude of the angle between $-\mu\beta$ and the
negative real axis, at $|\mu\beta| = 1$.   Figure 8 illustrates a phase margin of
about 50 degrees.   That is, if the points on the curve where $|\mu\beta| = 1.0$
were swung toward the negative real axis by about 50 degrees they would
coincide with the point $-1,0$, and the servo would be unstable.

The type of transient response obtained with reasonable gain and phase
margins is indicated in Fig. 9, which shows the response of the illustrative
servo system to an input step.   The initial overshoot is about 25%, and
the oscillation damps out very quickly.   For the general case, (6) may be
rewritten in the form

$$F_2 = \left[\frac{-\mu\beta}{1 - \mu\beta}\right] \cdot \frac{F_1}{-\beta}. \tag{8}$$

The relation $F_2 = F_1/-\beta$ may be regarded as the desired one, with the
bracketed factor acting as the inevitable modifier.   Then if the quantity

---

[9] With more complicated systems it may not be obvious whether or not the plot en-
circles $-1, 0$.   A simple test employs a vector with its origin at the $-1, 0$ point and its
tip on the curve.   If the vector undergoes zero net rotation as it traces along the curve
from $\omega = 0$ to $\omega = \infty$, the curve does not encircle the critical point.

[10] In some servo systems a *decrease* in amplifier gain also may cause instability.   Such
systems are still covered by the polar plot criterion of stability, and are commonly called
"Nyquist stable," or "conditionally stable."

$-\mu\beta$ exhibits gain and phase margins of the order of 10 db and 50 degrees respectively, the transient response of the modifying factor to a step function will be well-damped and generally not overshoot more than about 25%. If the gain margin is sufficient, the phase margin usually will be the dominant factor in determining the size of the initial overshoot. The required phase margin for critical damping depends upon the exact shape of $\mu\beta(j\omega)$, but in general is about 60 degrees. The gain margin needed in a particular design will depend upon the expected variability of the loop transmission. Radar tracking loops should usually have gain margins of the order of 15
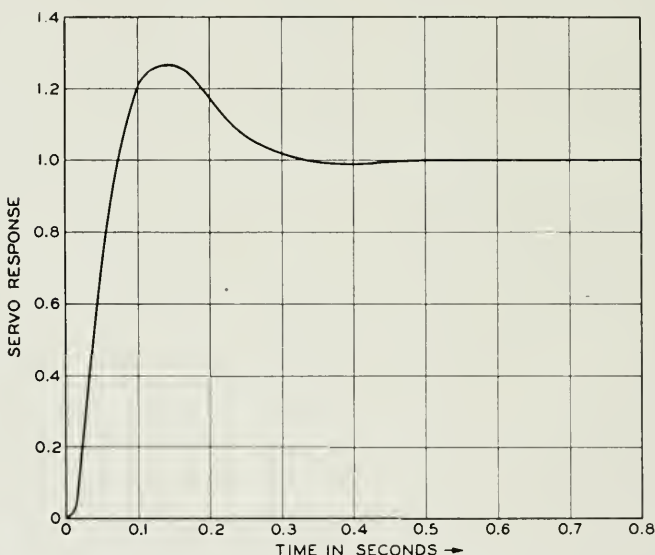

Fig. 9—Transient response of illustrative servo system.

db or more because of the large number of factors which may cause the loop gain to vary.

While the polar diagram gives a clear picture of stability considerations, it is usually more convenient for design purposes to plot the gain and phase of $-\mu\beta$ as separate curves on a logarithmic frequency scale, for positive values of ω. This is illustrated in Fig. 10, for the sample servo system. Under two commonly met conditions, the requirement for single loop[11] stability on this type of plot is simply that the absolute value of phase shift be less than 180 degrees at zero db gain ($|\mu\beta| = 1$). The conditions are that the connective polarity be such as to make $-\mu\beta$ positive when the

[11] Again, multiple loop systems may be included if all subsidiary loops are individually stable.

network phase shifts vanish, and that the gain curve cross zero db at only one frequency.[12]

An advantage of this logarithmic diagram is that commonly encountered forms of $|\mu\beta|$ vary as $\omega^{\pm n}$ for intermediate or asymptotic frequency regions, and thus plot as corresponding straight line segments. From (7.4) it may be seen that the illustrative form of $|\mu\beta|$ behaves, in turn, as $200\omega^{-1}$, $200\omega^{-2}$, $20\omega^{-1}$, and $1.6 \times 10^{8}\omega^{-4}$, as ω is increased. These asymptotic lines are drawn in lightly in Fig. 10, the actual gain describing smooth transi-
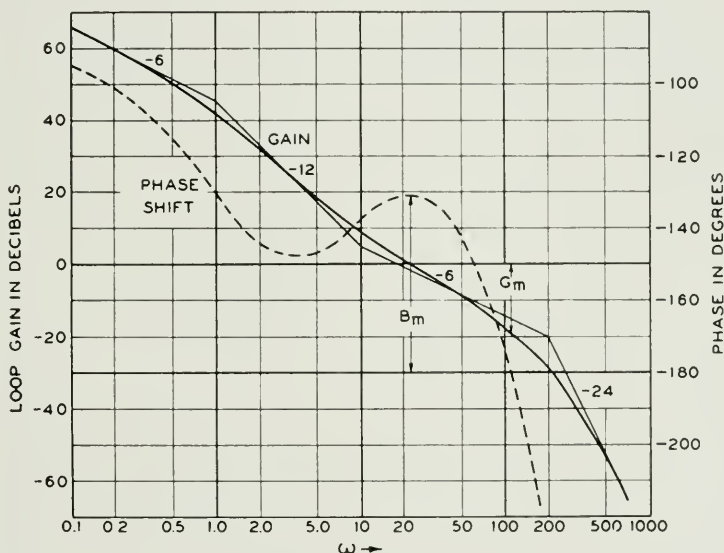

Fig. 10—Loop characteristic of illustrative servo system.

tions between adjacent asymptotes. Since the logarithmic slope of $\omega^{\pm k}$ is $\pm 6k$ db/octave,[13] the successive asymptotic slopes in Fig. 10 are -6, -12, -6, and -24 db/octave. The junctures of adjacent asymptotes occur at values of ω of 1, 10, and 200. These juncture frequencies are called "corner" frequencies, and may be seen from (7.2) to coincide with the real constants or "roots" added to $j\omega$ in each of the three factors in parentheses. The corner associated with the factor $200/j\omega$ is of course at $\omega = 0$. The corner of the last, or cubed factor is a multiple one, joining two asymptotes differing in slope by 18 db/octave. From a knowledge of such corner

[12] This discussion assumes that $\mu\beta$ has a low-pass configuration; that is, only the high frequency cut-off is considered. If $\mu\beta$ has a ow-frequency cut-off also, then a corresponding requirement identical with the above must be added.
[13] An octave is taken to be a 2:1 span in frequency, and 6 db is of course very closely equivalent to a 2:1 increase in $|\mu\beta|$.

frequencies, and the fact that the actual gain curve lies 3 db from an isolated simple asymptotic corner, the gain curve can usually be drawn in without further computation.[14]   The phase curve also is easily constructed by adding up the elementary phase curves associated with the various corners.   As may be seen from (7.3), these component phase curves all will have the same shape on a logarithmic frequency plot, merely being shifted along the frequency scale.   The phase contributed by each simple corner will be $\pm$ 45 degrees at the corner frequency, the sign depending upon whether the associated root appears in the numerator or in the denominator.

It is an extremely important fact that the very requirement of stability imposes an unambiguous interrelationship between the gain and the phase shift of most types of transfer characteristic!   By the general mathematical methods leading to the previously discussed stability criteria, Bode[15] has shown that this is true for the broad class of network structures commonly used in feedback loop design.   That is, if either the transfer gain or phase shift is specified *at all frequencies*, the attendant phase or gain can be computed without further information.   This class of networks is called *minimum phase*.   Any stable structure composed of lumped circuit elements will have a transfer characteristic of the minimum phase type, provided it does not include an all-pass section.[16]   All-pass characteristics are seldom used in the design of feedback loops, since their inclusion in the loop always reduces the stability margins achievable with a given high-frequency cut-off.   Thus the unique interrelationship between phase and gain may be assumed for the loop characteristic $-\mu\beta$ in single-loop feedback systems. The nature of this relationship is discussed in detail by Bode.   Briefly, the phase shift at any frequency ω_c is proportional to a weighted average of the gain slope in db/octave, over the entire logarithmic frequency scale. The weighting factor sharply emphasizes gain slopes in the immediate vicinity of ω_c , while the contributions of gain slopes at remote frequencies are reduced in proportion to the logarithmic frequency span from the particular frequency ω_c .[17]   For transfer characteristics of the form $\omega^{\pm k}$, having a constant gain slope of $\pm 6k$ db/octave,[18] the associated phase shift is also constant and equal to $\pm 90k$ degrees.   For transfer functions which behave approximately as $\omega^{\pm k}$ over a finite frequency span, the phase shift

---

[14] The corner frequency concept is less useful if the roots are complex.   However a great many servo systems are so constructed that $\mu\beta$ has only real roots.

[15] Loc. cit.   Also see "Relations between attenuation and phase in feedback amplifier design," by H. W. Bode, *B. S. T. J.*, July 1940, p. 421.

[16] An all-pass section is one which has constant gain but varying phase shift versus frequency, and is usually composed of a lattice, bridged T, or other bridge type circuit.

[17] About 60% of the area under this weighting function lies between $\omega = 0.5$ ω_c and $\omega = 2$ ω_c , 80% between 0.25 ω_c and 4 ω_c .

[18] That is, for transfer characteristics whose *absolute magnitude* is given by $\omega^{\pm k}\cdots$.

of $\pm 90k$ degrees is approached more and more closely as the length of span is increased.

This may be observed qualitatively from the transfer characteristic of Fig. 10. For $\omega \ll 1$, the gain slope is -6 db/octave, and the phase shift approaches -90 degrees. For $1 < \omega < 10$, the average gain slope is about -10 db/octave, and the phase shift near $\omega = 3$ is -148 degrees (instead of $-90 \times 10/6 = -150$ degrees). As ω increases toward 200, the phase shift increases rapidly due to the asymptotic slope of -24 db/octave, finally approaching -360 degrees ($-90 \times 24/6$) for $\omega \gg 200$.

Foreknowledge of the inevitable gain-phase relationship is of great value to the servo designer, in making clear the comparatively small class of realizable gain-phase combinations and thus averting attempts at non-physical designs. For example the design use of too-rapidly falling loop gain characteristics in the region of the high-frequency gain cross-over (that is, near zero db loop gain) is not permissible because of the large negative phase shifts which must accompany the steep gain slopes. Another way of stating the advantage of an early realization of the gain-phase laws is to say that the designer is assured in advance that any paired gain and phase characteristics which he chooses within the basic restrictions will be achievable with stable physical networks.[19]

3.2 *Dynamic Error*

A servo system is usually designed to transmit some class of input functions with a required degree of fidelity. This class of functions may reduce substantially to one specific input signal whose time variation or whose frequency spectrum is known, or it may include a great variety of signals which have certain properties in common. In the latter case it is conceivable that definite limits may be placed upon the allowable amplitude ranges of the input signal and its various time derivatives, or certain limiting frequency spectrum characteristics may be specified for the input function.

Servomechanisms are subject to several types of transmission error. The systematic error, or dynamic error, is predictable from knowledge of the *noise-free* input signal and of the transfer frequency characteristic of the servo system. For simplicity, the discussion of error will be limited to the case where the output signal is desired to be a replica of the input, and where $\beta = -1$. Thus the loop transmission $\mu\beta$ becomes simply $-\mu$. The input-output relationship as given by (6) is therefore

$$F_2 = \frac{\mu}{1 + \mu} F_1, \tag{9}$$

---

[19] With some necessary reservations as to practicable dissipation constants and parasitic circuit constants.

where F_1 and F_2 are again typical sinusoidal components of the input and output respectively. Thus the corresponding *sinusoidal error component* may be written as

$$\Delta = F_1 - F_2 = \frac{F_1}{1 + \mu}. \tag{9.1}$$

The methods which may be used to determine the actual dynamic error $\Delta(t)$ from (9.1) depend both upon the nature of $f_1(t)$ and the type of information available about $f_1(t)$. If the input signal is a known periodic function, $\Delta(t)$ may be found by applying (9.1) for each sinusoidal component of the input and summing the resulting terms. If the input is non-periodic in character, then the error may be calculated from the Fourier integral expression

$$\Delta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{F_1(\omega)}{1 + \mu} e^{j\omega t} \, d\omega, \tag{10}$$

where $F_1(\omega)$ represents the continuous frequency spectrum of $f_1(t)$, as obtained from

$$F_1(\omega) = \int_{-\infty}^{\infty} f_1(t) e^{-j\omega t} \, dt. \tag{10.1}$$

The problems of calculating $F_1(\omega)$ from $f_1(t)$ and $\Delta(t)$ from $F_1(\omega)$ often may be avoided by consulting well-known tabular lists of paired time and frequency functions.[20]

Equation (9.1) may be used as a broad guide in selecting the type of μ characteristic best suited to a particular input signal. It has been mentioned previously that because of input noise and parasitic circuit elements, the servo transfer bandwidth usually should be kept as narrow as possible, consistent with dynamic error requirements. The transfer characteristic $\mu/(1 + \mu)$ will be closely equal to unity while $|\mu| \gg 1$, will rise slightly[21] in the region where $|\mu| \backsimeq 1$, and fall off as μ when μ is small compared with unity. The "cross-over" frequency, for which $|\mu| = 1$, may be taken as a rough measure of the transfer bandwidth. Thus, the requirement of minimizing the bandwidth may be restated as that of minimizing the cross-over frequency, while holding the dynamic error within specified limits. Reasoning in a general way, this requirement may be met by designing μ so that the amplitudes of the sinusoidal error components, as given by

[20] An excellent list is given by G. A. Campbell and R. M. Foster in a Bell System monograph "Fourier Integrals for Practical Application," September, 1931. A table of Laplace Transforms, which also may be used, is given by M. F. Gardner and J. L. Barnes in "Transients in Linear Systems," John Wiley and Sons Inc., 1942.
[21] Assuming a phase margin of the order of 60 degrees.

(9.1), are roughly constant with frequency over the servo band. *This demands that μ have somewhat the same frequency distribution as the input signal spectrum* (for $|\mu| \gg 1$). Because of stability requirements and complexity of the necessary apparatus, this rule can usually be followed over only a part of the servo frequency band, especially when the input signal spectrum falls off very rapidly with increasing frequency. However, even a rough adherence to this desired relation is usually of real worth in reducing the noise errors of the servo. An illustration of this will be given in a later section.

3.21 *Approximate Calculation of Dynamic Error*

Frequently the servo requirement is to transmit, with great accuracy, a type of signal whose frequency spectrum falls off very rapidly with increasing frequency. As may be seen from (9.1) this demands very large values of loop transmission μ at the lower frequencies where the input signal energy is concentrated, but permits a rapidly dropping loop transmission versus frequency commensurate with the falling amplitude spectrum of the input signal. Such a rapid reduction in loop gain is practicable while $|\mu| \gg 1$. However stability considerations force a more gradual gain reduction as the region of gain cross-over is approached. As a result, contributions to the servo error from this frequency region may be neglected compared with those from the lower frequencies. This suggests a series expansion of (9.1) in the form,

$$\Delta = [a_0 + a_1(j\omega) + a_2(j\omega)^2 + a_3(j\omega)^3 + \cdots] F_1, \tag{11}$$

where a_0, a_1, etc. are real constants.

Because of the assumed rapid drop in component amplitude F_1 with increasing frequency it is often unnecessary to take account of more than a few terms of the expansion.[22]

It is easy to show that (11) may be rewritten on a time basis to give the total dynamic error as

$$\Delta(t) = a_0 f_1(t) + a_1 \dot{f}_1(t) + a_2 \ddot{f}_1(t) + \cdots, \tag{12}$$

where $(\dot{\ }) = d(\ )/dt$. Thus the coefficient a_0 gives the error component proportional to input displacement. Similarly, a_1 and a_2 are the coefficients of the error components due to input velocity and input acceleration, respectively. For a great many motor-drive servo systems the loop transmission μ approaches infinity as $1/j\omega$ when ω approaches zero. This en-

---

[22] The series may be said to converge rapidly in a practical sense, for the following reason: For small values of ω the higher order terms are negligible. For values of ω sufficiently large that the high order terms may no longer be neglected the coefficient F_1 has become so small as to make the contribution of the entire series negligible.

sures that a_0 and thus the displacement error will be zero, leaving principally the velocity and acceleration errors to be considered.

The coefficients a_0, a_1, a_2, etc. may be calculated easily for any particular case. For illustration, the three common forms of μ characteristic shown in Fig. 11 will be examined. (As previously discussed, the designated forms of μ need hold only for $|\mu| \gg 1$.)
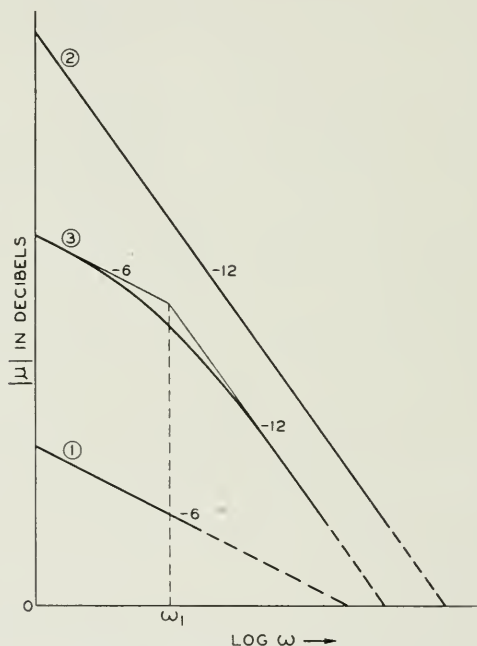

Fig. 11—Elementary μ forms.

Type 1. -6 db/octave, $\mu = \omega_0/j\omega$

The error expansion becomes,

$$\Delta(t) = \frac{1}{\omega_0}\dot{f}_1(t) - \frac{1}{\omega_0^2}\ddot{f}_1(t) + \cdots . \tag{12.1}$$

For the combination of high accuracy and rapidly converging input spectrum, the first term is the only one of importance. Thus this type of system has essentially a *velocity* error.

Type 2. -12 db/octave, $\mu = (\omega_0/j\omega)^2$

Here the error is

$$\Delta(t) = \frac{1}{\omega_0^2}\ddot{f}_1(t) - \frac{1}{\omega_0^4}\ddddot{f}_1(t) + \cdots . \tag{12.2}$$

Again for the rapidly converging case, this system will have principally an *acceleration* error.

Type 3. $-6, -12 \, db/octave, \, \mu = \omega_0\omega_1/j\omega(j\omega + \omega_1)$

This is perhaps the most commonly encountered characteristic in simple servos. The corresponding error expansion is

$$\Delta(t) = \frac{1}{\omega_0}\dot{f}_1(t) + \frac{1}{\omega_0\,\omega_1}\ddot{f}_1(t) - \frac{2}{\omega_0^2\,\omega_1}\dddot{f}_1(t) - \cdots, \qquad (\omega_0 \gg \omega_1). \qquad (12.3)$$

and the principal error for this type system thus is a combination of *velocity* and *acceleration* components. Either the velocity or the acceleration error component may be predominant, depending upon the various parameters.

3.3 *Noise Errors*

The typical sinusoidal component of servo error due to noise (unwanted signals or irregularities) in the input signal may be written as*

$$\Delta_n = \frac{\mu}{1 + \mu} N, \qquad (13)$$

where N represents the corresponding sinusoidal component of the input noise. If the noise signal $n(t)$ is known, the total noise error $\Delta_n(t)$ may be calculated from (13) in the ways described for the dynamic error. However, the noise input is seldom known in this sense, although certain outstanding components sometimes may be estimated and their effects evaluated. On the other hand the average disturbance due to random input noise, of the kind described as "thermal noise" in electrical circuits, may easily be calculated. This type of noise has constant amplitude versus frequency, and the total power in the output noise error may be found from

$$P_n = K \int_0^\infty \left| \frac{\mu}{1 + \mu} \right|^2 d\omega, \qquad (14)$$

where K is a constant dependent upon the input noise power.

Input noise also causes overloading of the power amplifier and overheating of the motor. These effects are aggravated by the falling transfer characteristic versus frequency of the motor, as seen from the following discussion. The servo transfer characteristic is maintained approximately at unity out to the cross-over frequency. However the transfer ratio of the motor, equation (3.1), will be falling at least at 6 db/octave, usually at 12 db/octave, at frequencies below this point.[23] Thus the transfer

---

* Again assuming $\beta = -1$.

[23] Assuming that the mechanical load impedance is a series combination of resistance and inertia.

characteristic (loop closed) from the servo input up to the motor and power amplifier must rise correspondingly with frequency, out to the cross-over point.   Again assuming input noise of the uniform amplitude versus frequency type, the total noise power at the motor input is therefore,

$$P_{nm} = K_1 \int_0^\infty \left| \frac{\mu}{1 + \mu} \right|^2 (\omega^2 + \omega_m^2)\omega^2 \, d\omega. \tag{15}$$

Again, ω_m is the reciprocal time-constant of the motor and K_1 is a proportionality constant.   If ω_m is less than about half the cross-over frequency, then the noise power at the motor input increases as the fifth power of the bandwidth of the servo transfer characteristic.[24]   Thus, if the input signal/noise ratio is small, this effect may be an important design consideration.

Still other servo errors may result from local extraneous signals or from coulomb and static frictional effects.   These error sources are in a somewhat different class from those discussed previously, in that they are more nearly under the designer's control.   That is, such extraneous signals and friction may be kept small by proper design and the residual friction effects further reduced by the use of local feedback.   In the absence of local feedback, the servo error resulting from frictional or other torque disturbances at the output shaft readily is found to be

$$\Delta_T = \frac{T}{S(j\omega)} \cdot \frac{1}{1 + \mu}. \tag{16}$$

$S(j\omega)$ is the actual stiffness (loop opened) of the output mesh, and T is the disturbing torque. T conceivably may represent static or coulomb friction, load-torque irregularities due to fluctuating running-friction, or wind torque.   Again assuming the mechanical impedance to be resistance and inertia in series, the mechanical stiffness is, from (2.2), $S(j\omega) = j\omega(R + j\omega J)$.   Thus the error is

$$\Delta_T = \frac{T}{j\omega(R + j\omega J)} \cdot \frac{1}{1 + \mu}, \tag{16.1}$$

and the apparent output stiffness (loop closed) is

$$S' = j\omega(R + j\omega J)\,(1 + \mu). \tag{16.2}$$

If T is taken as the static load torque, the resulting static error is found by setting $\omega = 0$ in (16.1).   Assuming that μ behaves as $\omega_0/j\omega$ when ω approaches zero, the static error is

$$\Delta_T = \frac{T}{\omega_0 R}, \tag{16.3}$$

[24] This assumes a constant functional form for the transfer characteristic.   However, the statement holds approximately, even with considerable variation in this form.

and the apparent low-frequency stiffness is $\omega_0 R$, being the ratio of the mechanical resistance to the velocity error coefficient. It may be noted that the static error will vanish if the loop transmission approaches infinity more rapidly than $1/\omega$ as ω approaches zero.

3.4 *Comparison of μ Characteristics for a Particular Input Signal*

In order to illustrate the advantages of shaping the loop characteristic for a particular input signal, a brief discussion will be given of the design of an automatic radar loop to track an airplane in azimuth over a constant linear-velocity course. The servo configuration is that given by Fig. 5b, θ_1 being the azimuth angle of the target and θ_2 the corresponding antenna angle. The lobing radar antenna has been assumed to take the place of the
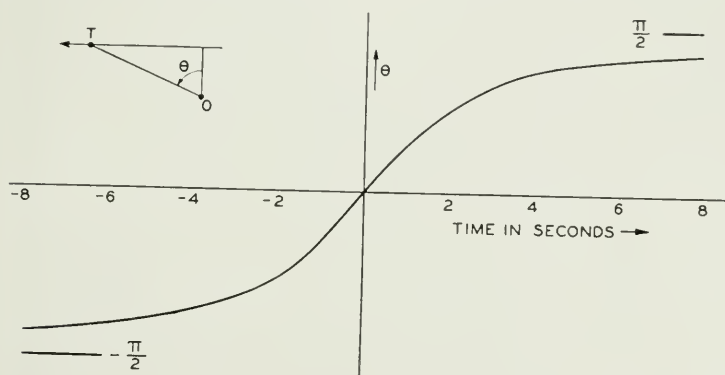

Fig. 12—Azimuth angle for constant linear velocity course.

synchro pair. Thus $\beta = -1$, and an error signal proportional to $\theta_1 - \theta_2$ is developed.[25]

Assuming a constant linear-velocity course having a maximum azimuth rate of 30 degrees/sec, the target azimuth angle is given by [26]

$$\theta_1(t) = \tan^{-1} .524t, \tag{17}$$

which is plotted in Fig. 12.

This course will develop a maximum azimuth acceleration $\ddot{\theta}_1$ of ± 10.3 degrees/sec² and a maximum $\dddot{\theta}_1$ of -16.4 degrees/sec³. The continuous frequency spectrum of $\theta_1(t)$ may be found from (10.1) to be

$$F_1(\omega) = \pi \frac{e^{-1.9|\omega|}}{j\omega}. \tag{18}$$

---

[25] Assuming a low elevation course.
[26] The azimuth angle has been so taken that zero azimuth is obtained at the point of nearest approach.

A logarithmic plot of $| F_1(\omega) |$ is shown in Fig. 13.*   It may be seen that the input signal spectrum falls at 6 db/octave for $\omega \ll 0.5$, at 12 db/octave for $\omega = 0.524$, and 30 db/octave at $\omega = 2.1$.

Assuming that the permissible dynamic error is 0.3 degree, a comparison will be made between the type 1 and type 3 loop characteristics of the previous section.   For the type 1 system, which will have essentially a pure velocity error, (12.1) shows the required value of ω_0 to be 30/0.3 or 100.
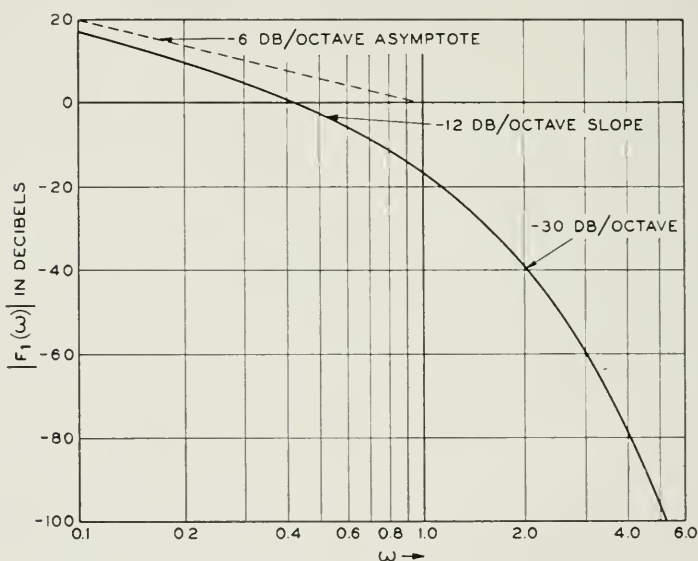

Fig. 13—Target frequency spectrum for constant velocity course.

Thus $\mu = 100/j\omega$.   Figure 14 shows a logarithmic plot of the corresponding $| \mu |$.   This characteristic departs rapidly from the shape of input signal spectrum given by Fig. 13, as ω is increased above 0.1.

The type 3 characteristic permits a considerably better match.   Choosing a compromise value for ω_1 of 0.1, (12.3) may be used to calculate the necessary value of ω_0 as 415.   Thus the loop transmission becomes $\mu = 41.5/j\omega(j\omega + 0.1)$.   Figure 14 shows a plot of the corresponding $| \mu |$, modified near the gain cross-over to satisfy the stability requirements. This curve is a considerably better average match for the target frequency-spectrum up to $\omega = 1$.   The resulting type 3 system has a predominant acceleration error as judged from the maximum velocity and acceleration errors of .072 degree and 0.25 degree respectively.

The total dynamic error curves for the constant-velocity course are given

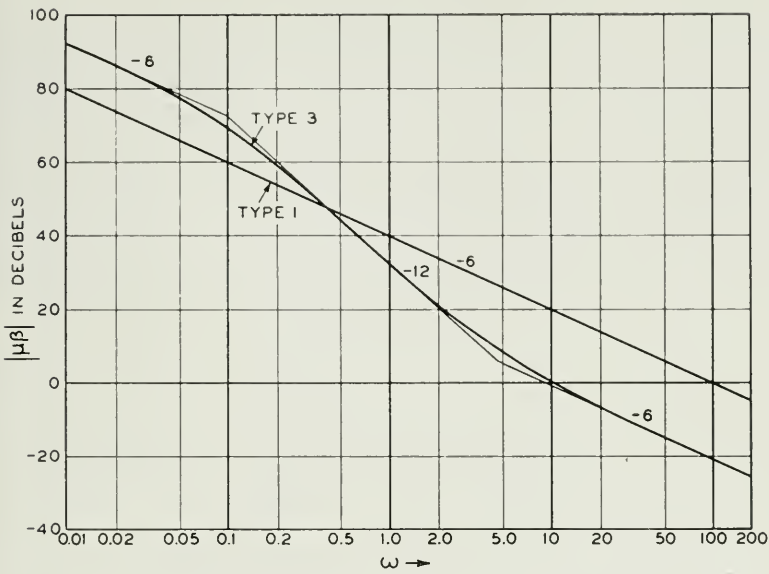* $| F_1(\omega) | = \pi$ has been taken as the zero db level.
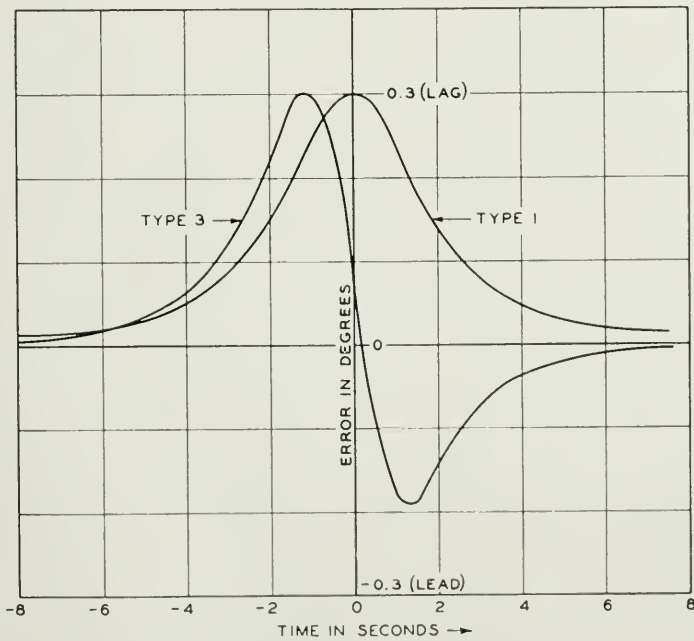
Fig. 14—Tracking loop characteristics.

Fig. 15—Tracking errors for constant velocity course.

in Fig. 15.   The velocity error of the type 1 system is always a lagging error
and is maximum at the point of nearest approach.   The type 3 composite
of velocity and acceleration errors is lagging over about the first half of the
course and leading for the second half, having lead and lag maxima at points
closely grouped about the point of nearest approach.

Although the two loop characteristics develop the same maximum dy-
namic error on the specified target course, their transient responses to an
input step differ widely, as may be seen from Fig. 16.   The rise time for the
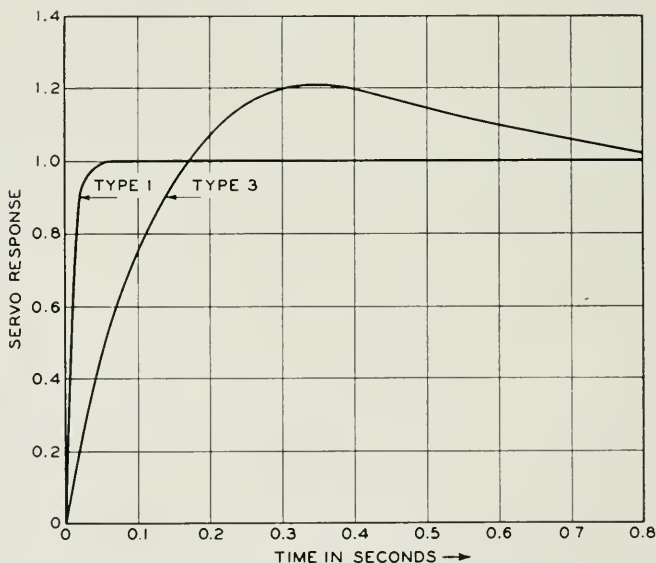type 1 loop is about .03 second compared with an initial rise in 0.17 second

Fig. 16—Transient response of tracking servos.

for the type 3 system.   Also, because of the overshoot the type 3 system
requires about 0.7 second to settle within 5% of the equilibrium value.

For a final comparison of the two systems the corresponding transfer
characteristics, $\mu/(1 + \mu)$, are plotted in Fig. 17 on arithmetic amplitude
and frequency scales.   It may be seen that the type 1 system is vulnerable
to noise and interfering signals over a far wider frequency range than the
type 3.   Again assuming uniform input noise versus frequency, (14) may
be used to show that the ratio of output noise power for the two systems is
about 7.5:1.

Thus the luxury of crisp transient response as obtained with the type 1
system may demand a heavy penalty in terms of output fluctuations due to
noise and other unwanted signal variations.   This is a clear illustration of

the necessity for designing the servo loop to match the type of input signal to be transmitted, particularly for radar tracking systems where the "unwanted variations" are ever present.

3.5 *Use of Local Feedback*

There are many examples of the use of local or subsidiary feedback in servo systems. The more common of these include feedback around vacuum tube power amplifiers to obtain improved linearity and impedance properties, and over-all feedback around amplifier and motor-drive systems to
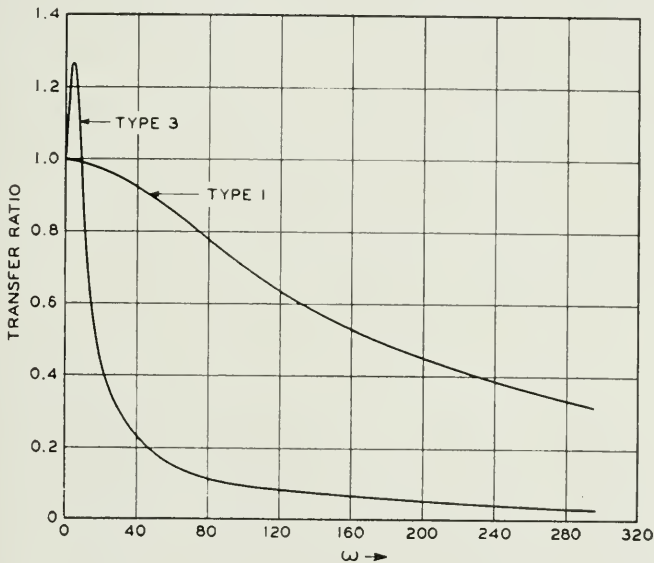
Fig. 17—Frequency response of tracking servos.

suppress frictional effects, increase output stiffness, and modify the inherent frequency characteristics of the basic components.[27] The tendency toward "β circuit dependency" as previously discussed also produces greater constancy of the stage transfer characteristics with time, temperature, etc.

Perhaps the simplest and most useful kind of local feedback is negative tachometer (velocity) feedback around motor-drive systems. This type of feedback widens the transfer frequency band of the drive system by reducing its time-constant, and increases the linear speed range of the motor. This may be illustrated by referring back to Fig. 7, which shows a typical tachometer loop. Assuming the transfer ratio of the amplifier to be a con-

[27] In a slightly different class are the servo systems used to provide automatic frequency and gain control in radio systems.

stant A, the transfer ratio of the motor and amplifier without feedback is, from (3.1),

$$\mu_T = \frac{\theta}{E} \text{ (loop open)} = \frac{\mu_0}{J} \frac{1}{j\omega(j\omega + \omega_m)}, \tag{19}$$

where the constant $\mu_0 = A\mu_t$. (To avoid confusion with primary loop quantities, the tachometer loop will be represented by the symbols μ_T and β_T, rather than μ and β.) The quantity ω_m was defined as the ratio $(R_m + R'_m)/J$ (see Fig. 2b), and is the reciprocal of the motor time constant. Replacing $(R_m + R'_m)$ by R for convenience, (19) may be rewritten as

$$\mu_T = \frac{\theta}{E} \text{ (loop open)} = \frac{\mu_0}{j\omega(R + j\omega J)}. \tag{19.1}$$

The transfer ratio of the tachometer is

$$\beta_T = \frac{E_\beta}{\theta} = -j\omega R_t,$$

and thus the loop transmission characteristic is

$$\mu_T \beta_T = -\frac{\mu_0 R_t}{R + j\omega J}. \tag{20}$$

For values of ω small compared with ω_m this loop transmission is constant and closely given by $\mu_T\beta_T(0) = -\mu_0 R_t/R$. When $\omega \gg \omega_m$, $\mu_T\beta_T$ approaches the form $-\mu_0 R_t/j\omega J$, and thus falls off at 6 db/octave. Consequently the maximum phase shift of the factor $-\mu_T\beta_T$ is -90 degrees, and no stability problem arises for the local tachometer loop.[28]

From (19.1) and (20), the over-all transfer ratio with feedback is

$$\frac{\theta}{E} \text{ (loop closed)} = \frac{\mu_T}{1 - \mu_T\beta_T},$$

$$= \frac{\mu_0}{j\omega(R + \mu_0 R_t + j\omega J)}. \tag{21}$$

Comparing (21) with (19.1), it may be seen that the sole effect of the tachometer feedback upon the over-all transfer ratio has been to add an apparent "ohmic" friction or mechanical resistance $\mu_0 R_t$ to the original value R. (It will be shown that this increase in apparent mechanical resistance also is effective in increasing the mechanical output impedance, although no power is dissipated in the added component $\mu_0 R_t$.)

---

[28] Actually, the effects of parasitic elements always modify this situation somewhat, especially if unusually high loop transmission is sought. However tachometer loops often require little or no stabilizing equalization.

Equation (21) also may be written as

$$\frac{\theta}{E} \text{ (loop closed)} = \frac{\mu_0}{J} \frac{1}{j\omega(j\omega + \omega'_m)}, \tag{21.1}$$

where $\omega'_m = (R + \mu_0 R_t)/J$ is the new corner frequency.

The change in over-all transfer ratio due to the tachometer feedback is shown in Fig. 18. The solid line diagrams A and B are the transfer gains without feedback and with feedback, respectively.[29] At low frequencies such that $\omega \ll \omega_m$, the feedback reduces the transfer ratio by the factor ω'_m/ω_m, the ratio of the two corner frequencies.[30] In order to restore this low-frequency loss in transmission, it is necessary to provide an added
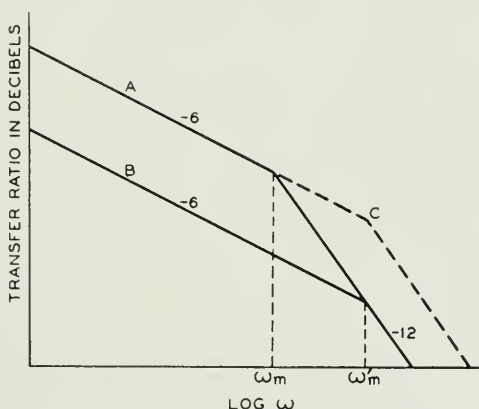

Fig. 18—Effect of tachometer feedback on motor characteristic.

amplification ω'_m/ω_m. If this is accomplished by increasing μ_0 and decreasing R_t so that the product $\mu_0 R_t$ remains constant,[31] the resulting transfer ratio will be that shown by the dotted lines C in Fig. 18. Comparing A and C, it may be seen that the net result of applying tachometer feedback and increasing the amplifier gain is to widen the transfer bandwidth by the factor ω'_m/ω_m. The required increase in amplification is the cost of widening the transfer bandwidth either by tachometer feedback or by non-feedback means, such as the use of a "forward-acting" equalizer in the amplifier. (However, such forward acting equalization fails to provide the increased over-all linearity and mechanical impedance obtained by the feedback method.) At frequencies sufficiently high that $\omega \gg \omega'_m$, the change in transfer ratio due to the feedback disappears, the mechanical inertia becoming the controlling element.

[29] The straight line asymptotes have been drawn instead of the actual gain curves.
[30] This is also the factor by which the feedback reduces the output speed obtained for a steady input voltage, neglecting circuit non-linearites and coulomb friction.
[31] This ensures a fixed loop transmission, and thus an unchanging value for ω'_m.

For ω small compared with ω_m', (21) becomes

$$\frac{\theta}{E} \text{ (loop closed)} \simeq \frac{\mu_0}{j\omega(R + \mu_0 R_t)}, \qquad (\omega \ll \omega_m').$$

If the tachometer feedback is substantial ($\omega_m' \gg \omega_m$), this may be further approximated as

$$\frac{\theta}{E} \text{ (loop closed)} \simeq \frac{1}{j\omega R_t}, \qquad \left(\begin{array}{c} \omega \ll \omega_m' \\ \omega_m' \gg \omega_m \end{array}\right). \qquad (21.2)$$

and the corner frequency becomes

$$\omega_m' \simeq \frac{\mu_0 R_t}{J}, \qquad (\omega_m' \gg \omega_m).$$

Thus for reasonably high feedback, the over-all transfer ratio (21.2) depends only upon the tachometer characteristic, being substantially independent of changes in the original mechanical resistance R or the amplifier-motor factor μ_0. The corner frequency ω_m' is similarly independent of changes in R, although still a direct function of μ_0. Thus the principal non-linearity of two-phase induction motors, namely variation in electrical damping with speed, is effectively suppressed by this type of local feedback, and systems employing such motors up to 80% of their synchronous speed may be designed on a linear basis.

The increase in mechanical impedance due to the feedback may be shown by assuming a torque disturbance T applied at the output shaft. Without feedback, the resulting speed disturbance is

$$\dot{\theta} \text{ (loop open)} = \frac{T}{Z_m} = \frac{T}{R + j\omega J}.$$

With feedback, the corresponding shaft speed disturbance becomes

$$\dot{\theta} \text{ (loop closed)} = \frac{T}{Z_m} \cdot \frac{1}{1 - \mu_T \beta_T},$$

$$= \frac{T}{R + \mu_0 R_t + j\omega J}.$$

Thus the apparent mechanical resistance, and therefore the protection against frictional torques, has been multiplied by a factor $(1 + \mu_0 R_t/R) = \omega_m'/\omega_m$. If the motor-drive system with tachometer feedback is employed in a simple follow-up system of the type of Fig. 5, equation (16.3) shows that the resulting low-frequency output-shaft stiffness will be $\omega_0(R + \mu_0 R_t)$ or $(\omega_m'/\omega_m)\omega_0 R.$[32] Therefore the output stiffness has

---

[32] The low-frequency loop transmission of the follow-up loop is again taken to be $\omega_0/j\omega$.

been increased by the factor ω'_m/ω_m over that obtained without the use of local feedback, assuming identical follow-up loop characteristics ($\mu\beta$) for the two cases. The ratio ω'_m/ω_m thus directly measures the feedback eduction of static and low-speed errors of the follow-up system due to torque disturbances. In practice the resulting increase in static accuracy may be of the order of 10 to 100 times.

3.6 *Error Reduction by Non-Feedback Means*

In situations where the noise associated with the input signal is small, it may be desirable to reduce the dynamic errors obtained with a given servo system by the use of forward-acting equalization external to the loop.

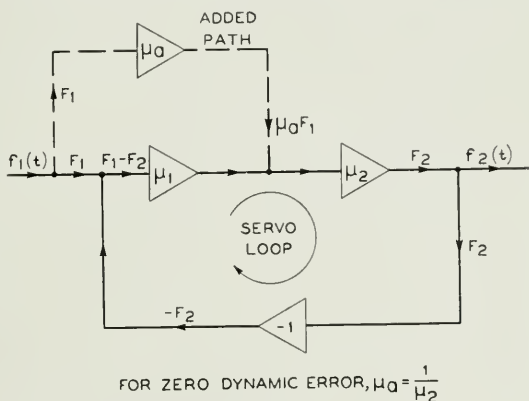

FOR ZERO DYNAMIC ERROR, $\mu_a = \dfrac{1}{\mu_2}$

Fig. 19—Forward-acting error compensation.

That is, the dynamic error characteristic may be computed, and the servo input or output modified by supplementary networks in such a fashion as to reduce the over-all error.

An illustrative arrangement, which is suitable when the input member is accessible,[33] is shown in Fig. 19. For convenience the servo is taken to be a simple follow-up system having $\beta = -1$. The μ circuit is shown divided into two parts, μ_1 and μ_2. Typically, μ_1 may be the transfer stiffness of a synchro pair (Fig. 3b), and μ_2 the transfer characteristic of a motor-drive system. The normal dynamic error component for such a loop, omitting the dotted line, has been shown to be $F_1/(1 + \mu)$. If an additional signal $\mu_a F_1$ is obtained from the input member and injected into the system as shown by the dotted line, then

$$F_2 = \frac{\mu}{1 + \mu}\,F_1 + \frac{\mu_a \mu_2}{1 + \mu}\,F_1,$$

$$= \frac{\mu + \mu_a \mu_2}{1 + \mu}\,F_1.$$

[33] This is not the case for a radar tracking loop, for instance.

Thus the over-all error becomes

$$F_1 - F_2 = \left(1 - \frac{\mu + \mu_a \mu_2}{1 + \mu}\right) F_1,$$

or

$$F_1 - F_2 = \frac{1 - \mu_a \mu_2}{1 + \mu} F_1. \tag{22}$$

If the added transmission path is so designed that

$$\mu_a = \frac{1}{\mu_2}, \tag{23}$$
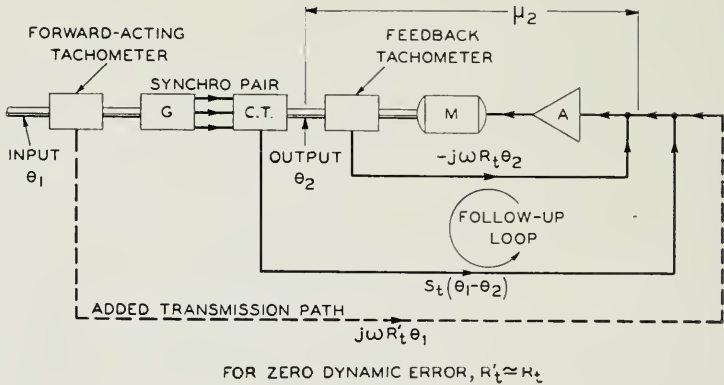

FOR ZERO DYNAMIC ERROR, $R'_t \approx R_t$

Fig. 20—Forward-acting tachometer system.

then $F_1 = F_2$, and the dynamic error vanishes. Thus the desired form of the added transmission depends only upon the μ_2 portion of the loop characteristic. It will not be possible to satisfy the condition given by (23) exactly, especially at the higher frequencies where noise enhancement and parasitic effects will become increasingly important. However, it is often possible to obtain the proper form for μ_a over the range of frequencies responsible for the bulk of the dynamic error. If μ_a has the proper frequency characteristic but is too large by 10%, for instance, it may be seen from (22) that there still remains a 10/1 increase in dynamic accuracy.

The foregoing method is especially applicable when μ_2 represents the transfer characteristic of a motor-drive system employing tachometer feedback, as shown in Fig. 20. Here the basic input-output comparison is obtained by means of the synchro pair, while a tachometer coupled to the input shaft provides the error-reducing signal. Thus the transmission μ_a is equal to $j\omega R'_t$, where R'_t is the tachometer transfer resistance. The expression for

μ_2 is given approximately by (21.2) as $1/j\omega R_t$. Thus, by (23), $R'_t \simeq R_t$ for substantial cancellation of the dynamic error (at frequencies small compared with ω'_m). That is, the output voltages of the two tachometers must closely annul each other when the input and output shafts are travelling at the same speed. Since the tachometers may be closely alike and excited from the same supply line, it is comparatively easy to keep their transfer ratios closely matched. In practice an error reduction of 20/1 is readily maintained by this method.

The error compensation scheme described above does not change the loop characteristic $\mu\beta$ of the basic servo loop, and thus does not create new stability problems. Its use to obtain high servo accuracy is desirable when the input noise is small and when a high loop gain is difficult to obtain because of parasitic elements or equipment complexities.

Abstracts of Technical Articles by Bell System Authors

Computation of Interfacial Angles, Interzonal Angles, and Clinographic Projection by Matrix Methods.[1] W. L. BOND. A way of setting up the general crystallographic axes *a*, *b*, *c* on unit orthogonal axes *x*, *y*, *z* is used to afford a matrix method of computing interfacial angles and zonal angles. It also affords a method of making clinographic projections.

A Current Distribution for Broadside Arrays which Optimizes the Relationship between Beam Width and Side-Lobe Level.[2] C. L. DOLPH. A one-parameter family of current distributions is derived for symmetric broadside arrays of equally spaced point sources energized in phase. For each value of the parameter, the corresponding current distribution gives rise to a pattern in which (1) all the side lobes are at the same level; and (2) the beam width to the first null is a minimum for all patterns arising from symmetric distributions of in-phase currents none of whose side lobes exceeds that level.

Design curves relating the value of the parameter to side-lobe level as well as the relative current values expressed as a function of side-lobe level are given for the cases of 8-, 12-, 16-, 20-, and 24-element linear arrays.

Paper Capacitors Containing Chlorinated Impregnants—Mechanism of Stabilization.[3] L. EGERTON and D. A. McLEAN. The stabilization of paper capacitors containing chlorinated aromatic impregnants with small quantities of organic additives is well established commercially. Although for practical reasons anthraquinone was chosen for initial commercial application, other quinones are also effective, as are the nitroaromatics, maleic anhydride, and sulfur. Evidence is given that the mechanism of stabilization consists of the formation of barrier films on the electrodes. These barrier films, which in certain cases may cover only the active points on the electrode surface, reduce the catalytic decomposition of the chlorinated impregnant by the electrode metal, prevent attack of the electrodes by liberated hydrogen chloride, and hinder electrolytic action. It appears likely that the film-forming properties of the stabilizers are dependent upon their oxidizing power. A secondary effect of stabilizers may be the formation of complexes with aluminum chloride to diminish the activity of the latter or change the nature of the reactions which it induces. Conductivity measurements in

[1] *American Mineralogist*, Vol. 31, pp. 31–42 (1946).
[2] *Proc. I.R.E. and Waves and Electrons*, June 1946.
[3] *Indus. and Engg. Chemistry*, May 1946.

HCl-saturated chlorinated diphenyl containing soluble additives are in line with known hydrogen-bonding tendencies of the additives. Compounds which are strong organic bases do not stabilize capacitors.

Quartz Crystals for Electrical Circuits.[4] R. A. HEISING. This book is a compendium of information, both theoretical and practical, on quartz crystal plates, their design and manufacture. It embodies the vast experience of the Bell Telephone Laboratories in research and in manufacture of quartz crystals. It originated from a series of lectures given by the members of the Laboratories technical staff who had carried out the early studies and developments in this field. By this means, engineers were trained for the immense expansion in crystal manufacture required to meet the demand of our military forces during the War. These lectures have been reorganized and rewritten, and are published together in this comprehensive book. Articles covering some of the various chapters have appeared in the Bell System Technical Journal.

The treatment covers in full the theory and practice of the preparation of quartz crystals, the instruments used, including new types developed for special purposes, the problems encountered in the various uses of quartz crystals, and the full details of the methods devised for their solution. The various processing chapters, dealing with cutting and grinding, plating and other topics of equal importance, include much information that appears for the first time in any book. The account of the manufacturing process is most complete. There are discussions of new practical methods of adjustment to frequency, of the new performance indicator, of a new type of crystal cut that operates at very low frequencies, and many new developments that represent notable advances in crystal technology.

Geometrical Characterizations of Some Families of Dynamical Trajectories.[5] L. A. MacColl. The chief problem considered in this paper is that of obtaining a set of geometrical properties which shall completely characterize the five-parameter family of trajectories of an electrified particle moving in an arbitrary static magnetic field. A solution of the problem is found in the form of a set of four principal and four subsidiary properties. A geometrical characterization, in the form of a set of two properties, is also given of the four-parameter family of trajectories of an electrified particle moving in an arbitrary static magnetic field with an arbitrarily prescribed value of the total energy. Various other properties of the families of curves are discussed, and the paper closes with a brief consideration of some analogous problems in which the particle moves in a fixed plane.

[4] Published by D. Van Nostrand Company, Inc., New York, N. Y., 1946.
[5] *Amer. Math. Soc. Trans.*, July 1946.

Comparison of Natural and Synthetic Hard Rubbers.[6] G. G. WINSPEAR, D. B. HERRMANN, F. S. MALM, and A. R. KEMP. GR-S, nitrile, and natural hard rubbers are compared as regards compounding, processing, vulcanization, and physical and dielectric properties. Natural rubber and GR-S compounds intermediate in sulfur content between hard and soft rubber also are compared. GR-S and nitrile rubber compositions suitable for commercial ebonite fabrication are described. Extensive breakdown of the basic copolymers has little effect on the physical properties of synthetic ebonites. The time required for the beginning of exothermic reaction in vulcanization is longer for GR-S than for natural rubber ebonites. Rockwell hardness is greater for GR-S. Some GR-S ebonites are penetrated to the same depth as natural ebonites, with a greater tendency toward instantaneous recovery. The two are similar in impact strength, but the ability to withstand a sharp bend is characteristic of natural ebonites alone. The latter are superior to GR-S ebonites in heat deformation below 60° C., but above this temperature the reverse is true and nitrile ebonites are superior to both. GR-S ebonites are more stable and nitrile ebonites less stable chemically than natural ebonites. GR-S ebonite dust as a filler increases brittleness. A diatomaceous earth improves the processing properties of GR-S hard rubbers. The adverse effect of ultraviolet light on surface resistivity is reduced when a GR-S hard rubber is filled with whiting. Natural and GR-S hard rubbers are alike in dielectric behavior.

Signal and Noise Levels in Magnetic Tape Recording.[7] D. E. WOOLDRIDGE. The primary object of the work described here was to determine what properties of the tape and associated magnetic elements are responsible for the noise and signal output levels of magnetic recordings and, if possible, to display in specific equations the pertinent relationships connecting noise and signal levels with the physical properties of the tape and polepieces. In the course of the study, methods appeared for decreasing the noise and increasing the useful signal reproduced from magnetic tape. These methods and some of the use that Bell Telephone Laboratories and Western Electric have made of them are mentioned in the discussion. While some of the work described in this paper has implications for more than one type of magnetic recording process, perpendicular recording on tape is the actual subject matter dealt with. In every case discussed, the record medium was 0.050 inch wide and 0.0022 inch thick. Except where otherwise noted, a chrome-steel tape was used at a speed of 16 inches per second.

[6] *Indus. and Engg. Chemistry,* July 1946.
[7] *Elec. Engg., Trans. Sec.,* June 1946.

Contributors to This Issue

WALLACE A. DEPP, B.S. in Electrical Engineering, University of Illinois, 1936; M.S. 1937. Bell Telephone Laboratories, 1937–. Mr. Depp has been primarily engaged in the development of gas filled tubes.

FREDERICK S. GOUCHER, A.B., Acadia University, 1909; A.B., Yale University, 1911; M.A. Yale, 1912; Ph.D. in physics, Columbia University, 1917; D.Sc. (Hon.), Acadia University, 1934. Engineering Department, Western Electric Co., 1917–18; Research, University College, London, 1919; Research Laboratories, General Electric Co., Ltd., England, 1919–26; Bell Telephone Laboratories, 1926–. In the Physical Research Department engaged in a fundamental study of contacts with reference to carbon microphones and switching apparatus, and during the war in the development of switching devices used in radar.

ROBERT E. GRAHAM, B.S. in E.E., Purdue University, 1937. Mr. Graham joined the technical staff of the Bell Telephone Laboratories in 1937, and has since been engaged in television research. During the war period he worked on radar and automatic tracking problems.

J. R. HAYNES, B.S. in physics, University of Kentucky, 1930. Bell Telephone Laboratories, 1930–. In the Physical Research Department engaged in contact studies and during the war the development of switching devices used in radar.

E. PETERSON, Cornell University, 1911–14; Brooklyn Polytechnic, E.E., 1917; Columbia University, M.A., 1923, Ph.D., 1926; Lecturer in Electrical Engineering, 1934–40. Electrical Testing Laboratories, 1915–17; Signal Corps, U. S. Army, 1917–19; Western Electric Company, Engineering Department, 1919–25; Bell Telephone Laboratories, 1925–. Dr. Peterson's work has dealt with non-linear circuits and circuit elements.

SNYDER C. RAPPLEYE, War Alumnus, Cornell University, 1919; New York Telephone Company, 1921–1923; Traffic Division of Operation and Engineering Department, American Telephone and Telegraph Company since 1923. Recently Mr. Rappleye has been working on intertoll trunk engineering methods.

E. J. Ryder, E.E., Polytechnic Institute of Brooklyn, 1935.   Engineering Department, Western Electric Co., 1922–25; Bell Telephone Laboratories, 1925–.   In the Physical Research Department engaged in contact studies and during the war the development of switching devices used in radar.